Vol. XXXV. Parts III and IV December 1948

BIOMETRIKA

FOUNDED BY

W. F. R. WELDON, FRANCIS GALTON AND KARL PEARSON

MANAGING EDITOR

E. S. PEARSON

ASSOCIATE EDITORS

M. G. KENDALL JOHN WISHART

in consultation with

HARALD CRAMÉR J. B. S. HALDANE

R. C. GEARY G. M. MORANT

MAJOR GREENWOOD

ISSUED BY

THE BIOMETRIKA OFFICE, UNIVERSITY COLLEGE, LONDON

PRINTED AT THE UNIVERSITY PRESS, CAMBRIDGE

Reprinted by offset litho 1953, 1963

[*Issued* 10 *January* 1949]

VOLUME XXXV, PARTS III AND IV DECEMBER 1948

SOME FURTHER NOTES ON THE USE OF MATRICES IN POPULATION MATHEMATICS

BY P. H. LESLIE

Bureau of Animal Population, Department of Zoological Field Studies, Oxford University

CONTENTS

		PAGE
1.	Introduction	213
2.	The stable female birth-rate	217
3.	The biological significance of the row vectors	219
4.	The total reproductive value of a population and the length of a vector	225
5.	The limited type of population growth	227
6.	The predator-prey relationship between two populations	238
	References	245

1. INTRODUCTION

The use of matrices in population mathematics has been discussed in a previous paper (Leslie, 1945), and some of the properties of the basic matrix representing a system of age-specific fertility and mortality rates have been described both there, and also in an earlier paper by Lewis (1942).† The purpose of the following notes is to enlarge on a few points left over from the earlier work, and in the later sections to extend the use of matrices and vectors to the case of the logistic type of population growth and to the predator-prey type of relationship between two or more populations.

In order to save a troublesome amount of cross-referring, it may perhaps be a convenience if the definitions and properties of the basic vectors and matrices are summarized here, and also if a brief account is given of the various transformations which are at one time or another used in the theoretical development. For fuller details reference may be made to the appropriate section of the original paper.

As before, for the sake of simplicity, the female population only will be considered, and the same unit of age will be adopted as that of time. If m to $m+1$ is the last age group in the complete life-table distribution defined by $L_x = \int_x^{x+1} l_x dx$ (taking $l_0 = 1$), and we put

$P_x\ (x = 0, 1, 2, \ldots, m-1) = L_{x+1}/L_x$ = the probability that a female aged x to $x+1$ at time t will be alive in the age group $x+1$ to $x+2$ at time $t+1$,

$F_x\ (x = 0, 1, 2, \ldots, m)$ = the number of daughters born in the interval t to $t+1$ per female alive aged x to $x+1$ at time t, who will be alive in the age group 0 to 1 at time $t+1$,

† At the time my original paper was published I was not aware that the same problem had already been investigated by Lewis (1942). This author establishes the form of the basic matrix and discusses a number of its properties, including the role of the dominant latent root and the form of the stable age distribution. He suggests that the rapidity with which an arbitrary age distribution settles down to the latter form will depend on the difference between the dominant and subdominant root of the characteristic equation, and he also discusses the type of matrix in which there is only a single non-zero element in the first row. It is clear, therefore, that unwittingly I covered a good deal of ground which had already been covered by him. I am indebted to Prof. M. S. Bartlett and Dr S. Vajda for this reference.

we are led to consider the square matrix M of order $m+1$ which has the F_x figures in the first row and the P_x figures in the subdiagonal immediately below the principal diagonal. For many purposes, however, it may not be necessary to deal with the matrix M as a whole. Thus, if $x = k$ is the last age group within which reproduction occurs, all the F_x figures for $x > k$ will be zero and the determinant $|M| = 0$. Partitioning the matrix symmetrically at this point the principal, non-singular, submatrix is

$$A = \begin{bmatrix} F_0 & F_1 & F_2 & F_3 & \dots & F_{k-1} & F_k \\ P_0 & . & . & . & \dots & . & . \\ . & P_1 & . & . & \dots & . & . \\ . & . & P_2 & . & \dots & . & . \\ \multicolumn{7}{c}{\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots} \\ . & . & . & . & \dots & P_{k-1} & . \end{bmatrix}.$$

As before an arbitrary age distribution will be written as the column vector ξ, different age distributions being distinguished by different subscripts. The number of elements composing a ξ vector may be either $m+1$ or $k+1$ depending on whether the particular age distribution considered is complete, or confined only to the pre-reproductive and reproductive age groups. Associated with each ξ_x there is a uniquely determined vector η_x, which in matrix notation is written as a row vector, the square of the length of the vector ξ_x being given by the scalar product $\eta_x \xi_x$. If the age distribution ξ_x is complete, consisting of $m+1$ elements, the last $m-k$ elements of the associated vector η_x will all be zero. Generally speaking, however, the post-reproductive age groups can be neglected, more particularly in the theoretical development, and unless otherwise stated it will be assumed that we are dealing with η and ξ vectors consisting of $k+1$ elements which are subject to the system of rates represented by the submatrix A.

It was shown in the previous paper (Leslie, 1945, § 5) that it is convenient for many purposes to pass to a new frame of reference, the vectors η and ξ and the matrix A undergoing the non-singular linear transformations

$$\eta = \phi H, \quad \xi = H^{-1}\psi, \quad B = HAH^{-1},$$

where H is a diagonal matrix with elements $(P_0 P_1 P_2 \dots P_{k-1})$, $(P_1 P_2 P_3 \dots P_{k-1})$, …, $(P_{k-2} P_{k-1})$, P_{k-1}, 1, which are derived entirely from the life table. (If the matrix M is the subject of the transformation instead of A, the matrix H may be suitably enlarged and will include all the P_x figures down to P_{m-1}.) It will be noted that in this collineatory transformation the square of the length of a vector is an invariant, and that the matrices A and B have the same characteristic equation and, therefore, the same latent roots.

The effect of this transformation on the elements of A is to replace the P_x figures in the principal subdiagonal by a series of units, and thus to reduce A to the rational canonical form. In biological terms it is equivalent to transforming the original population into one in which all the individuals live until the span of reproductive life is completed at the age of $x = k+1$. This imaginary type of population, with which in many ways it is more convenient to work, might be termed the canonical population.

When the relation between two column vectors is such that

$$B\psi_a = \lambda\psi_a,$$

where λ is a scalar, then ψ_a is termed a stable ψ appropriate to the matrix B. Similarly in the case of initial row vectors, if $\phi_a B = \lambda \phi_a$,

then ϕ_a is a stable ϕ appropriate to B.

It may be shown that corresponding to each distinct latent root λ_a of the characteristic equation of B, $|B - \lambda I| = 0$,

there is a pair of stable vectors ϕ_a and ψ_a which in the usual way may be normalized so that $\phi_a \psi_a = 1$. In the case when all the $k+1$ latent roots of B are distinct, the normalized stable ψ form a set of $k+1$ independent and mutually orthogonal vectors of unit length, and any arbitrary ψ_x may be expanded in terms of them, viz.

$$\psi_x = c_1\psi_1 + c_2\psi_2 + c_3\psi_3 + \ldots + c_{k+1}\psi_{k+1},$$

where the coefficients c_a may be either real or complex. Similarly the associated row vector ϕ_x can be expanded in terms of the stable ϕ,

$$\phi_x = \bar{c}_1\phi_1 + \bar{c}_2\phi_2 + \ldots + \bar{c}_{k+1}\phi_{k+1},$$

where $\bar{c}_a$ is the complex conjugate of c_a in the expansion of ψ_x. Similarly, by transforming back to the original co-ordinate system, any arbitrary ξ_x can be expanded in terms of the stable ξ and its associated vector η_x in terms of the stable η.

Since only one of the latent roots, and this the dominant one of the matrix B, is real and positive, only one of the stable ψ will consist of real and positive elements. It is this stable $\xi_1 = H^{-1}\psi_1$, associated with the dominant root λ_1, which is ordinarily referred to as the stable age distribution appropriate to a given set of age-specific fertility and mortality rates. The relation between the inherent rate of increase (r) and the dominant root of the matrix is given by $\log_e \lambda_1 = r$.

There is one further transformation of the matrix B which is of some theoretical importance. The expansion of an arbitrary ψ_x in terms of the normalized stable ψ may be written in matrix notation as $\psi_x = Qc_x$,

where the columns of the matrix Q consist of the stable ψ arranged from left to right in descending order of the moduli of the roots with which they are associated. In the same way the expansion of an arbitrary ϕ_x may be written

$$\phi_x = \bar{c}'_x U,$$

where $\bar{c}'_x$ is the transposed complex conjugate of the vector c_x, and the rows of the matrix U are formed by the stable ϕ arranged in a similar order from above down. Since the normalized stable vectors have the properties

$$\phi_a\psi_b \begin{cases} = 1 & (a = b), \\ = 0 & (a \neq b), \end{cases}$$

it follows that U and Q are reciprocal matrices ($UQ = I$). In this transformation to an orthogonal co-ordinate system the length of a vector remains an invariant and the matrix B becomes $UBQ = UHAH^{-1}Q = C$,

where C is a diagonal matrix whose elements are the latent roots of B (reduction to classical canonical form).

Since an arbitrary age distribution $\psi_x = H\xi_x$ must necessarily consist of real and positive elements, and since $\psi_x = Qc_x, \phi_x = \bar{c}'_x U$, we have $\phi_x = \psi'_x \overline{U}' U = \psi'_x G$, where G is a

symmetrical matrix of real elements. Thus, in terms of the original co-ordinate system, since H is a diagonal matrix unaltered by transposition,

$$\eta_x = \xi_x' HGH.$$

The matrix HGH, or G if the work is being carried out in terms of the canonical population, has the important property of converting a column vector into the associated row vector. The reciprocal relationship is given by

$$\xi_x = H^{-1}G^{-1}H^{-1}\eta_x',$$

where $G^{-1} = Q\bar{Q}'$. For further properties of the metric matrix G see the previous paper (Leslie, 1945, § 11).

It may perhaps be of interest if the actual values of some of these matrices are given for a simple numerical example, which will be used in some of the later sections in order to illustrate certain points. Although this example is purely a mathematical model bearing no relation to any known species, its properties are the same as those which might be observed for a population of living organisms considered in a small number of age groups, and for convenience biological terms will be used throughout in interpreting the results obtained with this matrix. Suppose, then, we have an entirely imaginary population which can be considered in four age groups, and let the life table or stationary age distribution be given by the L_x values forming the column vector $\{0{\cdot}9, 0{\cdot}7, 0{\cdot}5, 0{\cdot}3\}$. Further let the matrix

$$A = \begin{bmatrix} 0 & 45/7 & 18 & 18 \\ 7/9 & 0 & 0 & 0 \\ 0 & 5/7 & 0 & 0 \\ 0 & 0 & 3/5 & 0 \end{bmatrix}. \tag{1·1}$$

Then, since H is the diagonal matrix with elements $h_{11} = P_0P_1P_2$, $h_{22} = P_1P_2$, $h_{33} = P_2$, $h_{44} = 1$, we have

$$H = \begin{bmatrix} 1/3 & 0 & 0 & 0 \\ 0 & 3/7 & 0 & 0 \\ 0 & 0 & 3/5 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

and

$$HAH^{-1} = B = \begin{bmatrix} 0 & 5 & 10 & 6 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

The characteristic equation $|B - \lambda I| = 0$ is, when expanded in powers of λ,

$$\lambda^4 - 5\lambda^2 - 10\lambda - 6 = 0;$$

and the latent roots are therefore $\lambda_1 = 3$; $\lambda_2, \lambda_3 = -1 \pm i$; $\lambda_4 = -1$. In the transformation to the classical canonical form $UBQ = C$, the matrix

$$Q = \frac{1}{\sqrt{(68)}} \begin{bmatrix} 27 & 4{\cdot}9985 + 6{\cdot}4019i & 4{\cdot}9985 - 6{\cdot}4019i & \sqrt{(17)}\,i \\ 9 & 0{\cdot}7017 - 5{\cdot}7002i & 0{\cdot}7017 + 5{\cdot}7002i & -\sqrt{(17)}\,i \\ 3 & -3{\cdot}2010 + 2{\cdot}4992i & -3{\cdot}2010 - 2{\cdot}4992i & \sqrt{(17)}\,i \\ 1 & 2{\cdot}8501 + 0{\cdot}3509i & 2{\cdot}8501 - 0{\cdot}3509i & -\sqrt{(17)}\,i \end{bmatrix},$$

$$U = \frac{1}{\sqrt{(68)}} \begin{bmatrix} 1 & 3 & 4 & 2 \\ 2{\cdot}8501+0{\cdot}3509i & -3{\cdot}2010+2{\cdot}4992i & -13{\cdot}5488-7{\cdot}4545i & -7{\cdot}4977-9{\cdot}6029i \\ 2{\cdot}8501-0{\cdot}3509i & -3{\cdot}2010-2{\cdot}4992i & -13{\cdot}5488+7{\cdot}4545i & -7{\cdot}4977+9{\cdot}6029i \\ -\sqrt{(17)}\,i & \sqrt{(17)}\,i & 4.\sqrt{(17)}\,i & 6.\sqrt{(17)}\,i \end{bmatrix},$$

and

$$G = \bar{U}'U = \begin{bmatrix} 0{\cdot}5072 & -0{\cdot}4484 & -2{\cdot}1539 & -2{\cdot}1982 \\ -0{\cdot}4484 & 0{\cdot}8674 & 1{\cdot}9041 & 1{\cdot}5882 \\ -2{\cdot}1539 & 1{\cdot}9041 & 11{\cdot}2688 & 11{\cdot}2109 \\ -2{\cdot}1982 & 1{\cdot}5882 & 11{\cdot}2109 & 13{\cdot}4245 \end{bmatrix},$$

where in each case the elements have been rounded off to the fourth decimal place.

Since the ath row of the matrix U is the stable vector ϕ_a which is associated with the stable ψ_a vector given by the ath column of Q, it is possible to construct readily from their rows and columns the set of four matrices $S_a = \psi_a \phi_a$, which have the properties (Leslie, 1945, § 9)

$$S_a^2 = S_a, \quad S_a S_b = 0 \ (a \neq b), \quad \sum_a S_a = I.$$

If $f(B)$ is a polynomial of the matrix B, we have when the latent roots of the matrix are distinct,

$$f(B) = \sum_{a=1}^{k+1} f(\lambda_a) S_a.$$

Thus

$$B^t = \lambda_1^t S_1 + \lambda_2^t S_2 + \ldots + \lambda_{k+1}^t S_{k+1},$$

so that in the present example, when the matrix B is raised to a high power and λ_1^t is much greater than all the remaining λ_a^t,

$$B^t \propto \begin{bmatrix} 27 & 81 & 108 & 54 \\ 9 & 27 & 36 & 18 \\ 3 & 9 & 12 & 6 \\ 1 & 3 & 4 & 2 \end{bmatrix},$$

and hence, by transforming back to the original co-ordinate system,

$$H^{-1}B^tH = A^t \propto \begin{bmatrix} 81 & 312{\cdot}4283 & 583{\cdot}2 & 486 \\ 21 & 81{\cdot}0000 & 151{\cdot}2 & 126 \\ 5 & 19{\cdot}2857 & 36{\cdot}0 & 30 \\ 1 & 3{\cdot}8571 & 7{\cdot}2 & 6 \end{bmatrix}$$

for large values of t.

2. THE STABLE FEMALE BIRTH-RATE

Once the dominant latent root of the matrix has been found, there is one comparatively simple way of calculating the stable age distribution. Thus, working in terms of the canonical population and $m+1$ age groups, the stable ψ_1 appropriate to the root λ_1 may be taken proportional to the column vector $\{\lambda_1^m, \lambda_1^{m-1}, \ldots, \lambda_1, 1\}$, and by operating on this vector with the matrix H^{-1}, the stable age distribution ξ_1 can readily be obtained. The method which was used previously for calculating the stable female birth-rate was then to operate on this distribution with the maternal frequency figures† (m_x) and thus determine the total number

† The maternal frequency m_x is the mean number of live daughters born per unit of time to a female aged x to $x+1$. They are the figures tabulated in the usual type of fertility table and are not the same as the F_x figures forming the first row of the matrix.

of female births which might be expected per unit of time (Leslie, 1945, § 16). Although it seems likely that no very great error would be made in employing these methods, both the stable age distribution and the stable birth-rate can be defined rather more formally for the discontinuous case, and the appropriate equations can be derived for calculating them directly when the work is being carried out in terms of discrete age groups.

Consider at time t a stable age distribution $\xi(t)$ appropriate to the dominant latent root λ of the matrix M, and let n_x $(x = 0, 1, 2, \ldots, m)$ be the elements of this column vector. Then by the definition of a stable vector

$$\xi(t-x) = \lambda^{-x}\xi(t).$$

If $B(t)$ = the number of daughters born alive in the whole population in the interval of time t to $t+1$, it is easily seen that since n_x are the number of individuals alive aged x to $x+1$,

$$\lambda n_0 = L_0 B(t),$$
$$\lambda n_1 = P_0 n_0 = L_1 B(t-1),$$

and in general $\lambda n_x = L_x B(t-x)$.

If we put

π_x = the proportion of the stable population alive in the age group x to $x+1$,

and $N(t)$ = the total number of individuals alive in the stable population at time t;

$$\pi_x = \frac{L_x B(t-x)}{N(t+1)}.$$

Defining the birth-rate $\beta = B(t)/N(t)$,

we have in the case of the stable population,

$$B(t-x) = \beta N(t-x) = \beta N(t)\lambda^{-x},$$

so that

$$\pi_x = \beta L_x \lambda^{-(x+1)}, \tag{2·1}$$

an expression which defines the matrix stable age distribution. From this it follows, since

$$\sum_0^m \pi_x = 1,$$

that

$$\frac{1}{\beta} = \sum_{x=0}^{m} L_x \lambda^{-(x+1)}. \tag{2·2}$$

This argument for the case of discrete age classes is, of course, developed along lines similar to those followed by Lotka (e.g. 1939, p. 16) for the continuous case, where, if c_x is the proportion of the stable population aged between x and $x+dx$ and b the instantaneous birth-rate,

$$c_x = be^{-rx}l_x \quad \text{and} \quad \frac{1}{b} = \int_0^\infty e^{-rx} l_x dx. \tag{2·3}$$

The birth-rate β as defined by (2·2) is, however, a different type of birth-rate to that defined by (2·3). It is the total number of births taking place in the interval of time t to $t+1$ expressed per head of population at time t. If $D(t)$ is the number of deaths occurring in the same interval and $\delta = D(t)/N(t)$,

$$N(t+1) = N(t) + B(t) - D(t),$$

and thus, in the case of the stable population,

$$\lambda = 1 + \beta - \delta.$$

In order to express the relationship between β and b, we might consider that in the continuous case the number of births occurring during the interval of time t to $t+1$ will be given by

$$B(t) = bN(t)\int_0^1 e^{r\tau}\,d\tau,$$

whence

$$\beta = \frac{b}{r}(e^r - 1),$$

or, since $\log_e \lambda = r$,

$$b = \frac{\beta \log_e \lambda}{\lambda - 1}. \qquad (2{\cdot}4)$$

As an illustration of the comparative results obtained by applying these equations, we may take the same imaginary population of *Rattus norvegicus* as was used previously as a numerical example (Leslie, 1945). In the appendix to that paper it was shown that for the given system of fertility and mortality rates the value of r, estimated by the more usual methods of computation, was 0·44565, and that $b = 0{\cdot}51265$, this value of the birth-rate being obtained by the numerical integration of (2·3). When the system of rates was expressed in the form of a matrix of order 21×21, the dominant root was $\lambda_1 = 1{\cdot}56246$, or $r = 0{\cdot}44626$, and using equation (2·2) $\beta = 0{\cdot}64839$, and from (2·4) $b = 0{\cdot}5144$. The agreement between these estimates of the stable birth-rate is reasonably close and suggests that when we have already calculated the life table age distribution, which is so often the case, equations (2·2) and (2·4) of this section will provide an alternative method of calculating b, which would save a great deal of the tedious labour involved in the numerical integration of (2·3). Although theoretically it is necessary to consider the entire age span of the life table in applying these equations, this was not done in the present instance. In the numerical example given above the value of the rate of increase is so high that the post-reproductive age groups could be neglected without any very great error.

The stable birth-rate and death-rate of the transformed or canonical population (ψ_1 vector) are perhaps only of academic interest. In this connexion, however, there is a small point worth mentioning in order to correct a misstatement which was made in the previous paper. In a footnote (p. 208) it was there stated that 'in the transformed population the death-rate = 0'. Strictly speaking this would only be approximately true under certain conditions; for, if in the case of the stable canonical population

$$\lambda = 1 + \beta' - \delta',$$

where dashes are attached to the symbols in order to distinguish them from those used above, we have by putting $L_x = 1$ in (2·2) and carrying out the summation,

$$\beta' = \frac{\lambda^{m+1}(\lambda - 1)}{\lambda^{m+1} - 1},$$

and hence

$$\delta' = \frac{\lambda - 1}{\lambda^{m+1} - 1},$$

which will approach zero as λ^{m+1} becomes large. Actually in the numerical example given in the footnote referred to, the value of λ^{m+1} was sufficiently great for δ' to be taken as approximately zero without any very serious error being incurred.

3. THE BIOLOGICAL SIGNIFICANCE OF THE ROW VECTORS

The columns of the matrix M^t are a measure of the contributions made by each age group to the total population at time t. Thus, for example, if there were n_j individuals alive in the age group j to $j+1$ at $t = 0$, the number and age distribution of their living descendants and

survivors at time t could be found by multiplying the elements in the $(j+1)$th column of M^t by n_j, and hence their total contribution to the population at this time is given by n_j times the sum of these elements. It was shown previously (Leslie, 1945, § 4) that for values of $t \geqq m-k$, where $x = k$ is the last age group within which reproduction occurs, the last $m-k$ columns of M^t will consist only of zero elements, an expression of the obvious fact that individuals alive in the post-reproductive age groups contribute nothing to the population after they themselves are dead. From the point of view of the contributions made to the future population by the individual age groups, it is the submatrix A^t which is principally of interest. When t becomes very large, A^t can be taken as being proportional to the matrix

$$H^{-1}S_1H = H^{-1}\psi_1\phi_1H = \xi_1\eta_1,$$

and therefore the sums of the elements in the columns of A^t must be proportional to the row vector η_1. Since a population with an arbitrary age distribution tends ultimately to approach the stable form, provided that the system of age-specific fertility and mortality rates remains constant, it follows that the normalized row vector associated with the dominant latent root provides a measure of the relative contributions per head made to the stable population in the future by the individual age groups. Thus, supposing we have two arbitrary age distributions ξ_x and ξ_y, both subject to the same constant system of age-specific rates, the ratio between the total number of individuals in the two populations would, as time went on, tend to the figure

$$R = \frac{\eta_1\xi_x}{\eta_1\xi_y}.$$

If, instead of regarding ξ_x and ξ_y as two separate populations, we regard them as two components of an age distribution ξ_z, it is thus possible to estimate their relative contributions to the population in the future, subject to the condition that the system of rates represented by the matrix A remains constant.

If, in this expression for R, we put $\xi_y = \xi_1$, the normalized stable vector associated with the dominant root of the matrix, we may write

$$V = \eta_1\xi_x,$$

or, since the angle between two vectors ξ_x and ξ_y, of lengths x and y respectively, is

$$\cos\theta = \frac{\eta_y\xi_x}{yx}, \quad V = x\cos\theta_x,$$

where θ_x is the angle ξ_x makes with the stable vector ξ_1 of unit length. Thus, when ξ_x is the stable form of age distribution $(= c_1\xi_1)$, the quantity V is the same as the length of the vector ξ_x, since $\cos\theta_x = 1$, and when the population is not distributed as to age in the stable form, $0 < V < x$. The rate of increase of V with regard to time is $dV/dt = rV$, since $V(t) = \lambda_1^t V(0)$.

This quantity V appears to be essentially the same as that termed the total reproductive value of a population by Fisher (1930, p. 27). In discussing the equation

$$\int_0^\infty e^{-rx}l_xm_xdx = 1,$$

by means of which the inherent rate of increase r is usually calculated, Fisher points out the close analogy between a population increasing geometrically and the growth of capital invested at compound interest. Thus the birth of a child can be regarded as the loaning to him of a life and the birth of his offspring as a subsequent repayment of the debt. Then, 'a unit investment has an expectation of a return l_xm_xdx in the time interval dx, and the present value of this repayment, if r is the rate of interest, is $e^{-rx}l_xm_xdx$; consequently the

Malthusian parameter of population increase is the rate of interest at which the present value of births of offspring to be expected is equal to unity at the date of birth of their parent'. (In this quotation the original symbolism has been changed to that used here; Fisher writes m, the Malthusian parameter, instead of r, and the maternal frequency b_x instead of m_x.) Fisher then goes on to say that 'we may ask, not only about the newly born, but about persons of any chosen age, what is the present value of their future offspring; and if the present value is calculated at that rate determined as before, the question has a definite meaning—To what extent will persons of this age, on the average, contribute to this ancestry of future generations?' He then defines the reproductive value which can be assigned to a person aged x as

$$v_x = \frac{e^{rx}}{l_x}\int_x^\infty e^{-rt} l_t m_t \, dt.$$

Thus, by assigning to each of the n_x persons aged x the appropriate value v_x and summing over all age classes of a given age distribution, a figure which Fisher terms the total reproductive value of the population may be obtained. He also pointed out that this total reproductive value would increase or decrease according to the correct Malthusian rate r.

It was not difficult to show on an actual numerical example that the values of v_x were the same, apart from a scale factor, as the elements of the η_1 row vector after allowing for the fact that the latter refer to a population considered in discrete age groups, whereas the former refer to values of x which vary continuously; and it was evident that the calculation of the quantity V defined above was essentially the same as the calculation of Fisher's total reproductive value of the population.

There is, however, one important point in regard to the argument developed by Fisher which has been quoted. The present value of the repayment $l_x m_x dx$ is taken to be $e^{-rx} l_x m_x dx$, where r is the rate of interest. But, in the case of a population, this estimate of the present value would only be valid if the whole population were increasing at a rate r, and this would only be true when the stable form of age distribution was established. In other words, the reproductive value v_x assigned to a female aged x is the present value of her future daughters only when that female and her daughters are considered as members of a population with a stable age distribution. That this is so may be seen from a numerical example. Let us suppose we are given the age distribution

$$\xi_a = \{81, 21, 5, 1\},$$

which is a stable ξ appropriate to the dominant root $\lambda_1 = 3$ of the numerical matrix A (1·1) defined in the introduction. In one unit of time the population will be $A\xi_a = \lambda_1 \xi_a$ and these individuals will be either survivors or descendants of the original population. Each individual alive in the latter will contribute on the average so many living individuals to the population at $t = 1$, and we wish to assess the present value of that contribution. Consider first of all the solitary female alive in the last age group. In one unit's time this individual will be no longer alive, but she will have contributed $F_3 = 18$ living daughters to the population at that time. The present value of that contribution will therefore be $F_3/\lambda = 6$, and this is the present value which may be attached to each individual alive in this age group of a stable population at any given time. Passing to the five individuals in the next younger age group, $5P_2 = 3$ will be alive in the fourth age group at $t = 1$, and each of these three will be valued then at 6 or a total of 18. They will also have contributed $5F_2 = 90$ daughters. The present total value of the contribution made by these five individuals will be therefore

$$(90 + 18)/3 = 36.$$

or 7·2 per head. In the same way the 21 individuals in the second age group will each be valued at 3·85714 and the 81 in the first age group at 1 each. These values which have been determined in this way may be written as the row vector

$$\eta_1^* = [1,\ 3{\cdot}85714,\ 7{\cdot}2,\ 6],$$

where an asterisk is attached to the symbol in order to distinguish this vector from the true normalized form for this particular matrix, namely,

$$\eta_1 = \frac{1}{\sqrt{(68)}}[0{\cdot}3,\ 1{\cdot}28571,\ 2{\cdot}4,\ 2],$$

and it will be noted that $\eta_1^* = 3.\sqrt{(68)}\,\eta_1$.

It is clear from this example that this method of assessing the present value of the contribution made by each female aged x to $x+1$ to the population at time $t+1$ is equivalent to determining the present value of her future daughters, and that the valuation can only be carried out in this way when that female and her daughters are considered as members of a stable age distribution. Symbolically the equation which defines the elements y_x $(x = 0, 1, 2, \ldots, k)$ of the vector η^*, and which is equivalent to that given by Fisher for v_x in the continuous case, is

$$y_x = \frac{\sum_x^k \lambda^{-(x+1)} L_x F_x}{L_x \lambda^{-x}},$$

and by an obvious extension to the case of stable 'age distributions' consisting of complex or negative individuals, the stable η_a^* representing the present value of the 'contributions' made by each individual could be calculated similarly for each distinct latent root λ_a of a given matrix A. Moreover, it is evident in each case $y_x = 0$ for all values of $x > k$, the last age group in which reproduction occurs.

The use of these row vectors in the form η_a^* has, however, certain disadvantages, more particularly when it is necessary to compare the total present values of two stable age distributions which are each subject to a different system of rates of death and reproduction. It will be seen from the above equation defining v_x that if the maternal frequency is measured in terms of daughters, we must have in all cases $v_0 = 1$, since

$$\int_0^\infty e^{-rx} l_x m_x dx = 1 \quad \text{and} \quad l_0 = 1.$$

Similarly in the discrete case, the value of y_0 may be written, making use of the relationship $(P_0 P_1 P_2 \ldots P_x) = L_{x+1}/L_0$,

$$y_0 = \frac{F_0}{\lambda} + \frac{P_0 F_1}{\lambda^2} + \frac{P_0 P_1 F_2}{\lambda^3} + \ldots + \frac{(P_0 P_1 P_2 \ldots P_{k-1}) F_k}{\lambda^{k+1}},$$

which must be equal to unity, since from the characteristic equation of the matrix

$$\lambda^{k+1} - F_0\lambda^k - P_0 F_1 \lambda^{k-1} - \ldots - (P_0 P_1 \ldots P_{k-2}) F_{k-1}\lambda - (P_0 P_1 \ldots P_{k-1}) F_k = 0.$$

Thus, as exemplified in the numerical illustration given above, the vector η_1^* will always have its first element equal to unity and will in general differ from the normalized η_1 by some scalar factor. The vector η_1^* measures the total value of a stable population on a different scale, or in a different system of units, to those in which the present value is measured by the vector η_1. But the question of the respective units in which a number of such values are expressed might become of importance if two or more stable populations subject to different systems of rates were being compared. Suppose these rates are represented by a number of different

matrices $A_1, A_2, \ldots, A_n$, which will be assumed to be all of the same order. If the series of reproductive values for the individual age groups is taken as the row vector η_1^* appropriate to each of the given matrices, the first element of each vector will necessarily be unity as has been shown above. That the use of these vectors in this form for calculating the total present value may lead to unsatisfactory results for the comparison between two stable populations, can be seen from a simple example. Suppose each element of the numerical matrix A defined in the introduction (1·1), and which we will now call A_1, is divided by a factor of 3. The resulting matrix—say A_2—can then be taken as representing a new system of rates which has a dominant latent root $\lambda_1 = 1$. The stable age distribution $\xi_a = \{81, 21, 5, 1\}$ of A_1 is, however, also a stable ξ_a of A_2 appropriate to this root. If the stable η_1^* for the second matrix is calculated as before the elements will be the same as those given above for the original matrix. The total present value of the population represented by ξ_a would therefore be estimated at the same figure whichever of the two systems of rates it was subject to. If then these were two separate populations with rates A_1 and A_2, which happened to have identical age distributions, a comparison between them by means of the total values calculated in this way is not very informative. The easiest way out of this difficulty would be to use only the normalized η_1 associated with the dominant latent root of each matrix in calculating the total present value of a stable population for the purpose of comparing it with that of another. This procedure allows for any difference in what may be termed the respective scales of the two matrices. For this particular example, the normalized η_1 associated with the root $\lambda_1 = 1$ of the matrix A_2 is

$$\eta_{21} = \frac{0{\cdot}19245}{\sqrt{(68)}}[0{\cdot}\dot{3},\ 1{\cdot}28571,\ 2{\cdot}4,\ 2]$$
$$= 0{\cdot}19245\eta_{11},$$

where the initial of the two suffixes refers to the matrix with which the vector is associated. The total value of a population with an age distribution ξ_a would therefore be 8·2462 if it was subject to the system of rates represented by the matrix A_1, and 1·5870 when subject to A_2. Thus the use of the normalized row vectors instead of the form η_1^* leads to a different value being placed on each of the two populations corresponding to a difference in the systems of rates to which they are respectively exposed.

We may conclude, therefore, that in calculating the total value of a stable population it will in general be preferable to use the normalized stable row vector η_1 and not the form η_1^*. The one form, however, can be readily transformed into the other. For, working in terms of age distributions confined to the prereproductive and reproductive age groups, if the elements of η_1^* are calculated by means of the above equation for y_x, the relationship between η_1^* and η_1 is given by

$$\eta_1 = (P_0 P_1 P_2 \ldots P_{k-1})\left\{\frac{df(\lambda)}{d\lambda}\right\}^{-1}\eta_1^*,$$

where $df(\lambda)/d\lambda$ is the characteristic equation of the matrix differentiated with respect to λ, in which the numerical value of the dominant root is inserted and the square root taken with a positive sign. Thus, for the numerical example which has been used previously in this section, the characteristic equation of the matrix A defined by (1·1) is

$$f(\lambda) = \lambda^4 - 5\lambda^2 - 10\lambda - 6,$$

and

$$\frac{df(\lambda)}{d\lambda} = 4\lambda^3 - 10\lambda - 10.$$

For $\lambda_1 = 3$, we have $$\left\{\frac{df(\lambda)}{d\lambda}\right\}^{\dagger} = \sqrt{(68)},$$

and, since $P_0P_1P_2 = \frac{1}{3}$ for this matrix,

$$\eta_1 = \frac{1}{3\sqrt{(68)}}\eta_1^*,$$

corresponding to the difference between these two vectors which was noted above. Although this procedure has been illustrated in terms of the dominant root of the matrix, it can be similarly carried out for any stable η_a^* appropriate to a latent root λ_a. Alternatively, the normalized row vectors may be readily calculated in terms of the canonical population and the matrix $B = HAH^{-1}$ by the methods described in the previous paper (Leslie, 1945, §§7 and 8), and transformed back again by means of the relationship $\eta = \phi H$.

If Fisher's total reproductive value of a population is written in terms of vectors as the scalar $$V = \eta_1\xi_x = x\cos\theta_x,$$

it follows, as was pointed out earlier in this section, that when the population represented by the vector ξ_x is of the stable form of age distribution, we have $V = x$, the length of ξ_x. The total reproductive value, or the total present value, of a stable population is therefore given by the length of the vector representing the age distribution of the population. Now any population of individuals with a stable form of age distribution ξ_a can be represented as a multiple $c_1\xi_1$ of the normalized stable ξ associated with the dominant root of the matrix, and its associated vector η_a as a multiple $\bar{c}_1\eta_1$ of the normalized η_1, the square of the length of ξ_a being given by $\eta_a\xi_a$. We may thus regard the vector $\eta_a = \bar{c}_1\eta_1$, which is associated with the vector $\xi_a = c_1\xi_1$, as the representative of the population in terms of the individual present values according to age, just as the vector ξ_a is the representative of the population in terms of numbers according to age. Although we have been here considering only the total present value of a population of real positive individuals distributed as to age in the stable form, which must necessarily involve only one of the stable η or ξ for a given matrix, there is little difficulty from the mathematical point of view in considering 'populations' consisting of negative or complex individuals, and we may extend the arguments used for the real case so as to include all the stable vectors for the matrix. Thus, the length of any stable vector, ξ_a say, which fulfils the condition $A\xi_a = \lambda_a\xi_a$, can be regarded as the total present value of the 'population' represented in terms of numbers by ξ_a and in terms of individual present values by its associated vector η_a.

Since any arbitrary age distribution of real individuals ξ_x can be regarded as the sum of one or more mutually orthogonal stable ξ, viz.

$$\xi_x = c_1\xi_1 + c_2\xi_2 + \ldots + c_{k+1}\xi_{k+1},$$

and its associated vector η_x similarly as the sum of a number of associated stable η

$$\eta_x = \bar{c}_1\eta_1 + \bar{c}_2\eta_2 + \bar{c}_3\eta_3 + \ldots + \bar{c}_{k+1}\eta_{k+1},$$

and since the total present value of each of the component stable vectors is given by the length of that vector, namely $\sqrt{(\bar{c}_a c_a)}$, the total present value of the resultant ξ_x will be given by $\sqrt{(\sum \bar{c}_a c_a)}$, which is the length of the vector ξ_x.

The row vectors which were originally introduced into this theoretical discussion solely for mathematical reasons are thus not entirely without interest from the biological point of view. The uniquely determined vector η_x which was assumed to be associated with each ξ_x is a measure of the present value of the contribution made to future generations by an

individual aged x to $x+1$ when that individual is considered as a member of a population with an age distribution ξ_x. The row vectors appear to form a more generalized system of weights or values which we attach to an individual aged x to $x+1$ than the reproductive values v_x defined by Fisher. The latter are represented by a single member of this class of vectors, though one of particular importance owing to its association with the dominant root of the matrix.

Finally there is one further row vector which is very easily calculated for a given system of age-specific fertility and mortality rates, and which on occasion may be useful in studying the comparative fertility of different populations. The net reproduction rate,

$$R_0 = \int_0^\infty l_x m_x dx,$$

in addition to its usual meaning, may also be defined as the expected number of daughters which will be born on the average by a female now aged 0 during the remainder of her lifetime. It is in fact a figure which is analogous to the expectation of life at birth, only in terms of future daughters. Now, in addition to the newly born, we may also enquire what this expected number of daughters will be in the case of a female alive at any age x. Clearly this figure is given by,

$$u_x = \frac{1}{l_x}\int_x^\infty l_x m_x dx,$$

with $u_0 = R_0$. Similarly, in the discrete case, we may consider an η row vector of which the elements z_x $(x = 0, 1, 2, \ldots, k)$ are

$$z_x = \frac{1}{L_x}\sum_x^k L_x F_x,$$

and it will be found that this is merely a multiple of the η_1^* vector appropriate to the dominant root $\lambda_1 = 1$ of the matrix for a stationary population which is obtained by dividing each of the F_x figures in the first row of the matrix A by the net reproduction rate.

4. THE TOTAL REPRODUCTIVE VALUE OF A POPULATION AND THE LENGTH OF A VECTOR

It appears from the foregoing discussion that the elements of the normalized row vector η_1 can be regarded from two slightly different points of view. On the one hand they provide a measure of the relative contributions per head made by each age group to the stable population in the future, and this property arises from the fact that the sums of the columns of the matrix A^t can be taken as proportional to the elements of this vector when t becomes very large. On the other hand this vector is also associated with the column vector ξ_1 representing the stable age distribution appropriate to a given matrix, and in this sense its elements are a measure of the present value of the contribution made to future generations by an individual aged x to $x+1$ when that individual is considered as a member of a population with a stable age distribution. This difference is of importance in making any practical use of Fisher's total reproductive value of a population, which is defined here as $V = \eta_1 \xi_x$, where ξ_x is an arbitrary age distribution.

Thus, if we have two populations ξ_x and ξ_y both of which are subject to the same system of rates A, or alternatively if ξ_x and ξ_y are two subdivisions of one population subject to A, we can calculate for each the total reproductive values V_x and V_y, and determine the ratio $R = V_x/V_y$. This quantity, as was shown at the beginning of the previous section, is the ratio at time t, when t becomes very great, of the total number of individuals in the two populations which at $t = 0$ had the age distributions ξ_x and ξ_y. But the quantity R cannot be interpreted in this way when the two populations are not subject to the same system of rates.

Again, if a population happens to have a stable form of age distribution ξ_a, then $V = \eta_1 \xi_a = a$, the length of the vector ξ_a and this figure represents the total present value of the stable population ξ_a. But, apart from the case when an arbitrary ξ_x is of the stable form, it is difficult to define the meaning of V simply by itself in any precise biological terms. From the mathematical point of view, when an arbitrary ξ_x is expanded in terms of the stable ξ, and

$$\xi_x = c_1\xi_1 + c_2\xi_2 + \ldots + c_{k+1}\xi_{k+1},$$

we have

$$\eta_x \xi_x = \sum_{a=1}^{k+1} \bar{c}_a c_a,$$

which is the same thing as x^2, the square of the length of the vector ξ_x. Then it can be seen that since $V = \eta_1 \xi_x = c_1 = x \cos \theta_x$, the calculation of Fisher's total reproductive value is essentially the determination of one component of a set of mutually orthogonal sums of squares which together make up the total sum of squares represented by x^2. Thus $V^2 = c_1^2$ which is the first term in $\eta_x \xi_x = \sum_a \bar{c}_a c_a$, since c_1 is necessarily a real positive number.

The two methods of valuation which have been mentioned here are the calculation of the length of the vector ξ_x representing the age distribution of the population, and the calculation of the total reproductive value V. Which of these two figures is the more important from the point of view of assessing the state of a population subject to a given system of fertility and mortality rates is a matter for discussion and further investigation. Certainly the total reproductive value V is a figure which is the more easily determined. It requires only a knowledge of the row vector η_1 associated with the dominant root of the matrix representing the given system of rates to which the population is subject. On the other hand the calculation of the length of the vector ξ_x, is much more complicated. For, in order to arrive at the associated vector $\eta_x = \xi_x' HGH$, it is necessary to know the numerical values of the elements of the matrix G, and hence HGH, which in turn cannot be computed unless all the latent roots of the matrix A are known. Thus, purely from the practical point of view, the calculation of the total reproductive value $V = \eta_1 \xi_x$ offers a number of advantages and, within the limitations set out above, this figure may prove useful in comparing one population with another.

It is perhaps worth mentioning in passing one further type of problem. If the length of the vector ξ_x is regarded as the present value of the population when it is subject to a particular system of fertility and mortality rates, it may be of interest on occasion to consider the maximum or minimum of the quadratic form $\xi' HGH\xi$ given one or more restrictive conditions. Thus, for example, we might consider the problem of determining the column vector ξ_s which would give rise to the minimum total value when the sum of its elements was equal to a number N. If n_x $(x = 0, 1, 2, \ldots, k)$ are the elements of ξ_s and the symbol $\{1\}$ represents a column vector of $(k+1)$ units, we have, after differentiating with respect to the n_x and introducing a Lagrange multiplier λ,

$$HGH\xi_s - \lambda\{1\} = 0,$$

$$\Sigma n_x = N,$$

a set of $(k+2)$ equations for determining the values of n_x which will make the length of the vector ξ_s a minimum subject to the restrictive condition imposed. It will be seen from these equations that the solution of this problem is equivalent to that of determining the column vector ξ_s which will have all the elements of its associated row vector η_s the same value. Thus, by reversing the process, and starting with an arbitrary row vector of $(k+1)$ units, it

follows that the required column vector is proportional to the sums of the columns of the matrix $H^{-1}G^{-1}H^{-1}$. As an example of the type of vector which has the minimum value, the solution of these equations in the case of the simple 4×4 matrix given in the introduction was for $N = 108$, $\xi_s = \{84{\cdot}5686, 17{\cdot}4054, 4{\cdot}5059, 1{\cdot}5201\}$,

whereas the stable population of 108 individuals was

$$\xi_1 = \{81, 21, 5, 1\}.$$

This problem has been considered here in terms of the vector of shortest length, without imposing the full restrictive conditions which strictly speaking would be necessary when considering a population of living individuals, namely that the elements n_x of the column vector are positive integers with $\Sigma n_x = N$. But the vector ξ_s in this example consists of positive elements and may be taken as representing, in the case of this numerical system, the type of proportionate age distribution which would give rise to the minimum value. Actually the difference between the two distributions ξ_s and ξ_1 is not very marked in this example. The square of the length of the stable vector is 68, while that of the vector of shortest length is 64·4. But that this difference between the total values does correspond to a difference between the properties of the two age distributions may be seen by operating on each of them with the matrix A and determining the total number of individuals in the two populations at successive intervals of time. The numbers in the population which starts with an age distribution ξ_s will always be lower than those in the population starting with the stable form ξ_1, until ultimately there would be about 5·3 % fewer individuals in the former than in the latter.

5. The limited type of population growth

Hitherto it has been assumed that the system of age-specific fertility and mortality rates represented by the matrix A remains constant, and that therefore the population increases geometrically to an unlimited extent at a rate $dN/dt = rN$, when the stable age distribution is established. The next case which is usually considered in population mathematics is that of the logistic population, where the rate of increase in numbers is defined by the differential equation

$$\frac{dN}{dt} = (r - aN)\,N,$$

r and a being constants > 0, from which the well-known result follows that such a population will approach asymptotically an upper limit to the numbers given by $K = r/a$, according to the equation

$$N = \frac{K}{1 + Ce^{-rt}}.$$

It is therefore of interest to consider in terms of matrices and vectors the type of population growth in which the system of rates is dependent on the number of individuals present in the population at a given time.

Suppose that the system of rates to which a population is exposed when no limitations are placed upon the growth in numbers is represented by the matrix A with a dominant latent root λ_1. This might be called the optimum system of rates for the particular species or genetic stock. When the population is increasing in a limited environment let us suppose that at time t there is an age distribution $\xi(t)$ consisting of a total number $N(t)$ of individuals, and that at this time the elements of A are altered so that we have a new matrix A_t with a dominant latent root $\lambda_1/q(t)$, where $q(t)$ is dependent on $N(t)$. Then the age distribution of

the population at time $t+1$ will be given by $A_t\xi(t) = \xi(t+1)$, and the process can be obviously extended so that at time $t+1$ we have a matrix A_{t+1} with a dominant root $\lambda_1/q(t+1)$, $q(t+1)$ depending on $N(t+1)$, and so on. At each integral value of t, therefore, the original inherent rate of increase $r = \log_e \lambda_1$ will in general change to a new rate $r' = \log_e(\lambda_1/q)$, where q is some function of N, the number of individuals present in the population.

The changes which are thus assumed to occur in the optimum age-specific rates of fertility and mortality represented by the matrix A might take place in an innumerable variety of different ways. But, from the theoretical point of view, there are two extreme cases which are particularly of interest; on the one hand, when the decrease in the optimum rate of increase is due to a lowered degree of fertility, while the age-specific death-rates remain the same: and on the other when it is due to an increased rate of mortality and fertility remains constant. Even under these simplified conditions it is necessary to make some assumption as to the way in which the rates are actually affected, and in order to define the problem in concrete terms, it will be assumed here that the changes which occur either in the degree of fertility or in that of mortality are due to the operation of a factor which is independent of age. In addition one further type of change in the rates of fertility, involving a factor which increases geometrically with age, will be mentioned in passing. For simplicity the two main cases will be considered separately.

(*a*) *Mortality affected by a factor independent of age, fertility remaining constant*

If l_x and m_x are respectively the life table and fertility table for a population living under optimum conditions where no limitations are placed upon the growth in numbers, the inherent rate of increase (r) of the population is defined by

$$\int_0^\infty e^{-rx} l_x m_x dx = 1,$$

and the stable age distribution (c_x) and the stable birth-rate (b) by

$$c_x = be^{-rx}l_x, \quad \frac{l}{b} = \int_0^\infty e^{-rx} l_x dx.$$

If now a force of mortality (γ) which is independent of age is superimposed on the original force of mortality (μ_x), represented by the optimum life table l_x, the new life table l'_x will be given by

$$\frac{1}{l'_x}\frac{dl'_x}{dx} = -(\gamma+\mu_x) \quad \text{or} \quad l'_x = e^{-\gamma x} l_x;$$

and, if the original fertility table remains unaltered, the new inherent rate of increase will be $r' = r-\gamma$. The stable age distribution (c'_x) and stable birth-rate (b') of the population when it is subject to this new life table will then be

$$c'_x = b'e^{-r'x}l'_x, \quad \frac{l}{b'} = \int_0^\infty e^{-r'x} l'_x dx;$$

and it follows, since $l'_x = e^{-\gamma x}l_x$ and $r' = r-\gamma$, that $1/b' = 1/b$ and $c'_x = c_x$. The imposition of a force of mortality independent of age on a given life table thus leaves the original stable age distribution and stable birth-rate unchanged.

Similarly in terms of matrices, if A is the matrix representing the age-specific rates of fertility and mortality for a population living under optimum conditions, we are led to consider the matrix $q^{-1}A$ in which each element of the original matrix A is divided by a scalar q. Approximately, in the discrete case, this is equivalent to imposing on the original

life table a force of mortality which is independent of age. Then in the reduction of $q^{-1}A$ to rational canonical form, if

$$H_q = \begin{bmatrix} (P_0P_1P_2\dots P_{k-1})q^{-k} & & & \dots & . & . & . \\ & (P_1P_2\dots P_{k-1})q^{-(k-1)} & & \dots & . & . & . \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ . & . & . & \dots & (P_{k-2}P_{k-1})q^{-2} & . & . \\ . & . & . & \dots & . & (P_{k-1})q^{-1} & . \\ . & . & . & \dots & . & . & 1 \end{bmatrix},$$

the first row of $B_q = H_q(q^{-1}A)H_q^{-1}$ is

$$F_0q^{-1}, \quad P_0F_1q^{-2}, \quad P_0P_1F_2q^{-3}, \quad \dots, \quad (P_0P_1P_2\dots P_{k-1})F_kq^{-(k+1)},$$

the remaining elements consisting in the usual way of a series of units in the principal subdiagonal.

The characteristic equation of the matrix B_q is

$$\lambda^{k+1} - F_0q^{-1}\lambda^k - P_0F_1q^{-2}\lambda^{k-1} - \dots - (P_0P_1\dots P_{k-2})F_{k-1}q^{-k}\lambda - (P_0P_1\dots P_{k-1})F_kq^{-(k+1)} = 0$$

while that of the original matrix $B = HAH^{-1}$ is obtained by putting $q = 1$. Comparing these two equations term by term it will be seen that the latent roots of B_q are merely those of B each divided by the factor q. Thus, in terms of the canonical population, the stable age distribution appropriate to the dominant latent root λ_1/q of the matrix B_q may be taken as a multiple of the vector

$$\psi_1 = \left\{\left(\frac{\lambda_1}{q}\right)^k, \quad \left(\frac{\lambda_1}{q}\right)^{k-1}, \quad \dots, \quad \left(\frac{\lambda_1}{q}\right), \quad 1\right\}:$$

and since $\xi_1 = H_q^{-1}\psi_1$,

$$\xi_1 = \{(P_0P_1P_2\dots P_{k-1})^{-1}\lambda_1^k, (P_1P_2\dots P_{k-1})^{-1}\lambda_1^{k-1}, \dots, (P_{k-1})^{-1}\lambda_1, 1\},$$

which is the same as the $\xi_1 = H^{-1}\psi_1$ appropriate to the root λ_1 of the original matrix A. Moreover, since the time which it takes for an arbitrary ξ_x to approach the stable form of age distribution associated with the dominant root of the matrix will depend on the ratios of this root to the other roots of the matrix, as may be seen from the expansion of $\xi_x(t)$ at time t in terms of the stable ξ,

$$\xi_x(t) = c_1\lambda_1^t\xi_1 + c_2\lambda_2^t\xi_2 + \dots + c_{k+1}\lambda_{k+1}^t\xi_{k+1},$$

it follows that a population with any arbitrary form of age distribution which is subject to the matrix $q^{-1}A$ will approach the stable form at the same rate for all values of q. This result is of interest in the theoretical study of wild mammalian populations, since it might be assumed, at least as a first approximation, that any increase of mortality due to predation, hunger, etc., falling on some optimum system of age-specific death-rates could be represented by a factor which tended to be independent of age.

If then we consider at time t a population with an age distribution $\xi(t)$ which is subject to the system of rates represented by the matrix $q^{-1}A$, and we regard q as some function of N, the number of individuals present in the population at time t, we might put as a first approximation

$$q = \alpha + \beta N.$$

For the stationary state we must have $q = \lambda_1$, the dominant latent root of the matrix A,

and in addition, as N tends to zero, q must approach 1. When the dominant latent root of $q^{-1}A$ is equal to unity, the condition for a stationary population,

$$N = \frac{\lambda_1 - 1}{\beta} = K,$$

and therefore we may write $$q = 1 + \frac{(\lambda_1 - 1)N}{K}.$$

Then, assuming at time t there are $N(t)$ individuals distributed as to age in the stable form of distribution (ξ_a) for the matrix $q^{-1}A$, which distribution is the same for all values of q as has been shown above,

$$q^{-1}A\xi_a(t) = \xi_a(t+1) = \frac{\lambda_1 \xi_a(t)}{q},$$

or $$N(t+1) = \frac{\lambda_1 N(t)}{1 + \frac{\lambda_1 - 1}{K} N(t)},$$

and $$\frac{K - N(t+1)}{N(t+1)} = \lambda_1^{-1}\left\{\frac{K - N(t)}{N(t)}\right\},$$

which, as $\log_e \lambda_1 = r$, is the same thing as the logistic type of population growth,

$$N = \frac{K}{1 + Ce^{-rt}}.$$

Thus, when fertility remains constant and mortality is affected by a factor which is independent of age, this factor being regarded as a simple linear function of the numbers present in the population at time t, the total number of individuals in the population will increase according to the logistic form of population growth, provided that the age distribution of the population at $t = 0$ is the stable form appropriate to the dominant latent root of the matrix A. But, when this condition is not fulfilled, and the initial age distribution is not of the stable form, there may be quite considerable departures from the curve given by this actual logistic equation. The form of the curves representing the total number of individuals at successive intervals of time will, however, still tend to be S-shaped, and in some cases there is little doubt that a logistic type of equation could be fitted empirically to the data over a considerable portion of the total curve. The type of variation which might be expected in these growth curves owing to a departure from the stable form of age distribution is illustrated in the following simple examples.

Suppose that an entirely imaginary population, which can be considered in four age groups, is subject to the optimum system of rates of death and reproduction represented by the matrix defined originally in the introduction,

$$A = \begin{bmatrix} 0 & 6{\cdot}4286 & 18 & 18 \\ 0{\cdot}7778 & 0 & 0 & 0 \\ 0 & 0{\cdot}7143 & 0 & 0 \\ 0 & 0 & 0{\cdot}6000 & 0 \end{bmatrix},$$

which has a dominant root $\lambda_1 = 3$, or $r = 1{\cdot}09861$, and suppose that for the matrix $q^{-1}A$,

$$q = 1 + 0{\cdot}000185185N.$$

where N is the number of individuals in the population at integral values of time t. When

$q = 3$, the stationary state, $N = 10800$; and at $t = 0$ let there be 108 individuals present in the population. These conditions are fulfilled by the logistic equation

$$N = \frac{10800}{1 + 99e^{-1.098612t}}.$$

If at $t = 0$ we consider three different age distributions each consisting of 108 individuals and represented by the vectors

$$\xi_a = \{81, 21, 5, 1\}, \quad \xi_s = \{85, 17, 4, 2\}, \quad \xi_x = \{0, 0, 108, 0\},$$

where ξ_a is a stable age distribution of the matrix A, ξ_s the vector of shortest length given in the previous section and expressed to the nearest integer, and ξ_x a very skew form of age distribution in which all the individuals are concentrated in an age class for which fertility is high, the age distributions and therefore the total number in each population can be readily calculated by successive applications of the matrix $q^{-1}A$. The following are the results obtained in each case, together with the values of N calculated from the logistic equation

Values of N

t	From logistic	Initial age distribution		
		ξ_a	ξ_s	ξ_x
0	108·0	108	108	108
1	317·6	318	292	1970
2	900·0	901	844	1930
3	2314·3	2316	2215	6199
4	4860·1	4862	4660	8423
5	7673·7	7675	7540	9389
6	9508·7	9509	9433	10694
7	10332·3	10332	10298	10609
8	10639·5	10641	10628	10741
9	10745·9	10745	10742	10804
10	10781·9	10781	10780	10781

which are given in the first column. It will be seen that in the case of the stable age distribution ξ_a the values of N follow those calculated from the logistic equation, apart from small discrepancies at times in the last figure due to errors of rounding off. (The elements of the vector $\xi(t+1) = q^{-1}A\xi(t)$ were in each case expressed to the nearest whole number.) In the case of ξ_s, an age distribution which does not differ very greatly from the stable form, the numbers lie below those for the initial distribution ξ_a until $t = 10$, the stable age distribution being approximately established in this population round about $t = 7$; while for ξ_x the numbers are very erratic owing to the very skew form of the initial distribution leading to a very rapid increase in numbers during the early stages. The stable form of age distribution was approximately established in this last population at $t = 10$. It is evident from these examples that the initial form of age distribution may have a marked effect on the course of development followed by a population which inherently is increasing towards some upper limit according to the type of growth in numbers assumed here.

The initial number of individuals in these three examples is small relative to the upper limit of $K = 10800$, so that even a thoroughly skew form of distribution such as ξ_x has time

in which to approach the stable form of age distribution before the upper limit in numbers is achieved. Actually in the case of ξ_x the stable form is not established before $t = 10$, and a tendency to overshoot the upper limit will be noticed before that time. If the initial number of individuals had been chosen much greater relative to K this tendency would only have been emphasized. An extreme case would have been to assume that the initial number of individuals in each of the three examples was equal to 10800. Then it is evident that whereas the population represented by ξ_a, the stable form, would have remained constant at the same figure, those represented by ξ_s and ξ_x would vary on either side of the upper limit to begin with and would tend to approach the steady state by a series of damped oscillations as the stable age distribution was in the process of being established.

(*b*) *Fertility affected by a factor independent of age, mortality remaining constant*

This problem raises a number of difficulties not all of which have been satisfactorily resolved. But, before considering the main problem as defined here, namely when fertility is affected by a factor independent of age, there is another case which arises from the foregoing discussion, and which is perhaps worth mentioning. The canonical matrix

$$B_q = H_q(q^{-1}A)\,H_q^{-1}$$

defined above, is when written in full, to take a simple example of a 4×4 matrix,

$$B_q = \begin{bmatrix} F_0q^{-1} & P_0F_1q^{-2} & P_0P_1F_2q^{-3} & P_0P_1P_2F_3q^{-4} \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix};$$

and it can be seen that in addition to being the canonical form of $q^{-1}A$, B_q is also the canonical form of

$$A_q = \begin{bmatrix} F_0q^{-1} & F_1q^{-2} & F_2q^{-3} & F_3q^{-4} \\ P_0 & 0 & 0 & 0 \\ 0 & P_1 & 0 & 0 \\ 0 & 0 & P_2 & 0 \end{bmatrix},$$

the diagonal matrix H of the transformation HA_qH^{-1} having elements $h_{11} = P_0P_1P_2$, $h_{22} = P_1P_2$, $h_{33} = P_2$, $h_{44} = 1$. This matrix A_q can be regarded as representing some system of age-specific rates in which an original level of fertility included in the F_x figures has been affected by a factor which increases geometrically with age, and as before this factor q might be taken as being linearly related to N, the number of individuals present in the population at time t. But in contradistinction to the matrix $q^{-1}A$, the age distribution of a population subject to the matrix A_q will no longer remain stable. For suppose that in terms of the canonical population the stable age distribution associated with the dominant root λ_1/q of A_q is

$$\psi_1 = \left\{\left(\frac{\lambda_1}{q}\right)^k, \left(\frac{\lambda_1}{q}\right)^{k-1}, \ldots, \left(\frac{\lambda_1}{q}\right), 1\right\},$$

the transformation $\xi_1 = H^{-1}\psi_1$ gives

$$\xi_1 = \left\{(P_0P_1P_2 \ldots P_{k-1})^{-1}\left(\frac{\lambda_1}{q}\right)^k, (P_1P_2 \ldots P_{k-1})^{-1}\left(\frac{\lambda_1}{q}\right)^{k-1}, \ldots, P_{k-1}^{-1}\left(\frac{\lambda_1}{q}\right), 1\right\},$$

which is not the same as the ξ_1 associated with the dominant root λ_1 of the original matrix A.

If then at time t a population happened to have the stable form of age distribution appropriate to the matrix $A_q(t)$, it will not in general have the stable form of distribution at $t+1$ appropriate to $A_q(t+1)$, except in the case of the stationary population with $N = K$ and $\lambda_1/q = 1$, when the life table age distribution is established.

By extending this argument for the matrix A_q to the perfectly general case, it can be seen that the age structure of a population will be constantly changing when the degree of fertility is affected and the life table remains constant, until in the terminal stages of its growth the population approaches the stationary state. This is, of course, essentially the same type of changing age distribution as that shown to occur by Lotka (1931) in the case of a population growing in numbers according to the logistic law with a constant form of life table. A numerical example is given later of a population subject to the matrix A_q when q is taken as a simple linear function of N.

Although, biologically speaking, it is not impossible for fertility to be affected by a factor which increases geometrically with age and which depends on the number of individuals present in the population at a given time, it is perhaps of greater interest to consider the case in which the fractional decrease in fertility is the same at all ages. In other words, it is necessary to consider the matrix A_s, say, in which the elements in the first row of a matrix A representing the optimum rates of death and reproduction are each divided by a factor s, so that

$$A_s = \begin{bmatrix} F_0 s^{-1} & F_1 s^{-1} & F_2 s^{-1} & \dots & F_k s^{-1} \\ P_0 & . & . & \dots & . \\ . & P_1 & . & \dots & . \\ \dots & \dots & \dots & \dots & \dots \\ . & . & . & P_{k-1} & . \end{bmatrix}.$$

Now, if the $(x+1)$th element in the first row of the canonical form $B = HAH^{-1}$ is written as

$$(P_0 P_1 P_2 \dots P_{x-1} F_x) = f_x,$$

the characteristic equation of the original matrix A is

$$\lambda^{k+1} - \left(\sum_{x=0}^{k} f_x \lambda^{k-x} \right) = 0,$$

and that of the matrix A_s is $\lambda^{k+1} - \left(s^{-1} \sum_{x=0}^{k} f_x \lambda^{k-x} \right) = 0.$

If the real positive root of the first equation is λ_1, the real positive root of the second can be written as λ_1/q, and the inherent rate of increase of a population subject to the system of rates A_s will be $r' = \log_e(\lambda_1/q)$. Since we are considering as before the case when q is a function of N, say

$$q = 1 + \frac{\lambda_1 - 1}{K} N,$$

it is necessary, in order to solve the problem of a population in which fertility is affected by a factor independent of age, that s should be expressed as a function of q.

This point proved to be rather troublesome, and the following solution needs a much fuller investigation than it has received here. It depends on the relation between the first row of the canonical form $B = HAH^{-1}$ and the $L_x m_x$ column which was touched on in the previous paper (Leslie, 1945, § 6). It is evident that the division of the elements in the first row of the

matrix A or B by a scalar s is the same thing as dividing the maternal frequency figures (m_x) by the same quantity. The original net reproduction rate, $R_0 = \int_0^\infty l_x m_x dx$, will therefore become R_0/s. Now, in the solution of the equation

$$\int_0^\infty e^{-rx} l_x m_x dx = 1,$$

we have

$$\log_e R_0 = m_1 r - \frac{m_2}{2!} r^2 + \frac{m_3}{3!} r^3 - \frac{m_4 - 3m_2^2}{4!} r^4 + \ldots, \tag{5·1}$$

where

$$m_1 = \int_0^\infty x l_x m_x dx \Big/ \int_0^\infty l_x m_x dx,$$

and m_n $(n = 2, 3, 4, \ldots, n)$ is the nth moment about this mean. When the maternal frequency is divided by s the moments of the distribution will not be affected, but the value of r will change to a new value r', and

$$\log_e (R_0/s) = m_1 r' - \frac{m_2}{2!} r'^2 + \frac{m_3}{3!} r'^3 - \frac{m_4 - 3m_2^2}{4!} r'^4 + \ldots. \tag{5·2}$$

The moments are usually calculated by treating the $L_x m_x$ figures $\left(L_x = \int_x^{x+1} l_x dx\right)$ as a frequency distribution, the individual frequencies being regarded as centered at the midpoint of each age group. Alternatively they are sometimes calculated from $l_x m_x$, where l_x is the value of the usual life table function taken at each midpoint. When a system of rates is expressed in the form of a matrix the elements of the first row of the canonical form $B = HAH^{-1}$ are not the same as the $L_x m_x$ figures. But it was found (Leslie, 1945, § 6) that the sum of these elements was equal to the net reproduction rate and that if each element $(P_0 P_1 P_2 \ldots P_{x-1} F_x)$ was regarded as centered at the age of $x+1$, the mean and semiinvariants of the distribution were the same as those obtained from the $L_x m_x$ column.

These relationships suggested a possible way of relating s to q. If for the matrix A, with a dominant latent root $\lambda_1 = e^r$, the sum of the elements in the first row of $B = HAH^{-1}$ is equal to R_0, and if the dominant latent root of the matrix A_s is $\lambda_1/q = e^{r'}$, we might, as a first approximation, take only the first terms in each of the equations (5·1) and (5·2), and put

$$\log_e R_0 = m_1 r, \quad \log_e (R_0/s) = m_1 r',$$

and

$$\frac{r'}{r} \log_e R_0 = \log_e (R_0/s),$$

or, since

$$\frac{r'}{r} = 1 - \frac{\log_e q}{r},$$

$$\log_e s = \frac{\log_e R_0}{r} \log_e q. \tag{5·3}$$

For a greater degree of accuracy the first two terms could be taken as (5·1) and (5·2), viz.

$$\log_e R_0 = m_1 r - \frac{m_2}{2} r^2, \quad \log_e (R_0/s) = m_1 r' - \frac{m_2}{2} r'^2;$$

and

$$\frac{r'}{r}\left(1 + \frac{r - r'}{\frac{2m_1}{m_2} - r}\right) \log_e R_0 = \log_e (R_0/s).$$

From which, putting $\log_e q = w$ and $\frac{2m_1}{m_2} - r = c$,

$$\log_e s = -\frac{\log_e R_0}{rc}\{(r-c)w - w^2\}. \tag{5·4}$$

For a greater degree of accuracy still, further terms on the right-hand side of (5·1) and (5·2) could be included, though the algebra tends to become somewhat tedious. Presumably the number of terms which it would be necessary to include in any particular case would depend on the magnitude of r and upon the form of the distribution relating net fertility to age. Actually in the elementary numerical example which has been used here so far, equation (5·4) appears to be fairly accurate. Thus the characteristic equation of A with $\lambda_1 = 3$ and $R_0 = 21$ is

$$\lambda^4 - 5\lambda^2 - 10\lambda - 6 = 0.$$

Dividing these numerical coefficients by $s = 5$, for example,

$$\lambda^4 - \lambda^2 - 2\lambda - 1{\cdot}2 = 0,$$

of which the real positive root is $\lambda_1/q = 1{\cdot}63476$, or $q = 1{\cdot}83513$. The values of m_1 and m_2 were 3·04762 and 0·52154 respectively, and equation (5·4) was in common logarithms

$$\log s = 2{\cdot}48366 \log q + 0{\cdot}60263 (\log q)^2. \tag{5·5}$$

For $q = 1{\cdot}835$, the estimated value of s is 4·975, whereas the true value is $s = 5$. If s is estimated from equation (5·3) for $q = 1{\cdot}835$ the value is 5·379, so that the second degree equation in $\log q$ is an improvement on the first and gives a reasonably close approximation to s for values lying in this region. It will be noted that if $q = 3$, $s = 21$ from this second degree equation (5·5), as it should do.

In order to compare the operational effect of the matrix A_s with that already determined for $q^{-1}A$, two examples are given below for the initial age distributions

$$\xi_a = \{81, 21, 5, 1\}, \quad \xi_x = \{0, 0, 108, 0\},$$

ξ_a being the stable age distribution of 108 individuals for the matrix A, and ξ_x the same form of skew distribution used previously. As before, q was taken as

$$q = 1 + 0{\cdot}000185185N,$$

and the appropriate value of s at each stage was calculated by means of equation (5·5). In addition one example is given of the operation of the matrix A_q in which fertility is affected by a factor which increases geometrically with age, taking ξ_a as the initial distribution. The results were as follows:

Values of N

t	From logistic	Matrix A_q ξ_a	Matrix A_s	
			ξ_a	ξ_x
0	108·0	108	108	108
1	317·6	312	312	1915
2	900·0	867	867	1976
3	2314·3	2118	2115	5603
4	4860·1	4194	4120	7315
5	7673·7	6659	6393	8616
6	9508·7	8857	8362	10464
7	10332·3	10268	9696	10369
8	10639·5	10876	10384	10695
9	10745·9	10984	10673	10901
10	10781·9	10900	10766	10715

Comparing the two cases in which the initial distribution was of the stable form ξ_a with the figures derived from the logistic curve, it will be seen that in both cases the numbers of individuals are less than those for the logistic particularly in the early stages of development. Broadly speaking, however, all these three curves are similar in their general outlines, though there is an obvious tendency in the case of the matrix A_q for the population to overshoot the upper limit of $N = 10800$ in the later stages. Similarly, in the case of the initial distribution ξ_x and the matrix A_s, the course of events is not very different from that for the previous example with this distribution, when it was assumed that mortality was changing and fertility remained constant, though, again here, the numbers of individuals are less when fertility is changing and mortality remains the same. The chief difference between these examples and those given previously lies, of course, in the forms of the age distribution. When the matrix $q^{-1}A$ was assumed to be in operation, the ultimate age distribution to which all populations would tend, whatever their initial conditions and numbers might be, was

$$\xi = \{8100, 2100, 500, 100\};$$

whereas, both for the matrix A_q and A_s, the stationary age distribution of 10800 individuals is

$$\xi = \{4050, 3150, 2250, 1350\};$$

and throughout the whole course of development of each population an approach is being made to one or other of these very different distributions.

Although the two extreme cases of either fertility or mortality changing through the operation of a factor which is independent of age have been considered here separately, there should be little difficulty in extending the methods so as to include the case where both fertility and mortality are affected in varying degrees at the same time. Thus, we might consider the scalar q of the dominant latent root λ_1/q at time t as being the product, $q = uv$, of two factors, one of which, u say, represents an increase in mortality independent of age, and the other v represents the effect of a decrease in fertility at all ages by means of the factor s. Various possibilities then arise, depending on whether the ratio u/v was regarded as a constant, or as varying in some predetermined manner. However, these questions have not been gone into any further at present.

It will be noticed that the problem considered in this section of a growing population subject to a changing degree of fertility and a constant life table is not precisely the same as that discussed by Lotka (1931). In the first part of that paper Lotka showed how the birth-rate, death-rate, age distribution and inherent rate of increase of such a population would change when the total number of individuals in the population increased according to the logistic law. Here no assumption is made as to the way in which the number of individuals is increasing, but it is assumed that at equal intervals of time, which intervals in practice can be made as small as we please according to the degree of accuracy required, the inherent rate of increase of the population $r' = \log_e(\lambda_1/q)$ is dependent on the number of individuals (N) present at time t, and, as a first approximation, q has been taken as a linear function of N. The most important feature of this form of population growth is the marked effect which the initial age distribution and numbers have on the subsequent course of development of the population. Only in one case, namely when mortality is increased owing to the operation of a factor independent of age, fertility remaining constant, and when the initial age distribution is of the stable form appropriate to the matrix A, is the true logistic form of growth in numbers realized. However, the result of operating on a not too abnormal initial distribution with either of the matrices $q^{-1}A$, A_q or A_s is, broadly speaking, a very similar type

of S-shaped curve, if the initial numbers are small relative to the upper limit K, and in some cases there is little doubt that a logistic equation could be fitted empirically to such a series of points, more particularly when the figures for the total number of individuals are not available over the complete range of development of the population. But, in general, we shall have for a given matrix A and a given value of K in the equation $q = 1+(\lambda_1-1)N/K$, a family of S-shaped or partially S-shaped curves (or even the type of curve which descends towards the upper limit K), the differences between the individual members depending on the initial state of the population and on the way in which the decrease in the inherent rate of increase takes place, whether through a decrease in fertility, or an increase in mortality, or a combination in varying degrees of both factors. Among the more interesting features of this type of population growth is the possibility, under suitable initial conditions, of the total numbers in the population becoming greater than K and then of finally approaching the stationary state by means of a series of damped oscillations around this limit.

It is interesting to consider in the light of these results some of the population growth curves which have been published for one or other species of insect living alone in a limited environment (e.g. Chapman, 1928; Crombie, 1945). Certainly the initial age distribution of some of these populations must have been extremely skew, consisting as they did in many cases of only a small number, perhaps only a pair, of adults. It is a little difficult, on looking through the figures given in these various papers, to rid oneself of the impression that some of the curves may have been influenced, in part at least, by these rather extreme initial conditions. But at present this remains an impression and nothing more; it does suggest, however, that the part played by the initial age distribution is worth investigating further in these experimental populations.

Although the dominant latent root of the matrix operating between t and $t+1$ has been considered here only as a function of the number of individuals present at time t, there should be little difficulty in extending the argument so as to include the case when q is assumed to be a function not of $N(t)$ but of $N(t-a)$ where a is an integer, or even of an integral, $\int_0^t N\,dt$ say. This last would be equivalent to assuming that the growth of the population was defined by a type of integro-differential equation such as is introduced by Volterra in his development of population mathematics (e.g. Volterra, 1931, p. 141; Volterra & D'Ancona, 1935, p. 22). Moreover, there is another and more speculative approach which is not without interest. In all these various forms of population growth the inherent rate of increase is regarded as dependent on the total numbers and thus each individual is counted as being of the same value for all age distributions of which it is a member. In other words, the factor q is taken to be some function of the scalar $[1]\xi$, where $[1]$ is a row vector of units. Now, from the biological point of view, it is not unreasonable to suppose that the form of the age distribution may also be of importance. For a given value of N we might have two entirely different age distributions, one of which was composed largely of adult individuals and only a small number of young, and the other with these proportions reversed. The question naturally arises whether one is justified in assuming that both the populations are of equal value and that they both influence the system of rates to the same extent. The one with the larger proportion of adults might exert a greater degree of influence on the rate of increase owing, for instance, to a proportionately greater consumption of food, or an enhanced mutual interference between the individual members of the population. But this is at present purely

speculative, and so far as the writer is aware, there is no experimental evidence for the occurrence of such differential effects associated with the form of the age distribution when the populations are of the same size. As a possibility, however, it is of interest theoretically and it suggests that instead of counting all individuals as equal, some system of weighting the individual age classes would be required. A mathematical model which immediately comes to mind is that of a matrix whose dominant latent root is affected by the length of the vector on which it is operating; that is to say, it would be assumed that the inherent rate of increase was dependent on the present value of the population at a given time.

6. The predator-prey relationship between two populations

It is of interest to consider very briefly a simple type of predator-prey relationship between two species of which the one, S_1, is preyed upon by the other, S_2. If the matrix A_1 with a dominant latent root λ_1 represents the optimum system of rates for the prey and the matrix A_{1t} for this population at time t has a dominant root λ_1/q_1, we might regard the factor q_1 as a function of N_2, the number of the predatory species S_2, and write as a first approximation,

$$q_1 = 1 + \alpha_1 N_2, \tag{6·1}$$

where $\alpha_1 > 0$ is a constant. In the same way there will be some optimum system of rates A_2 for the species S_2, though in fact this system may never be realized in full save under exceptional circumstances, for instance when the prey are extremely numerous in comparison with the predator, and everything in the environment is favourable to the latter species. (From the biological point of view there must be some upper limit to the possible inherent rate of increase of which a particular species is capable. For instance, in the case of mammals, this limit will be determined in part by physiological factors, such as the length of the gestation period, the shortest interval between litters, the maximum average number of daughters per litter, the age at which breeding first starts, and so forth, as well as the form of life table under the most favourable circumstances.) Then at time t the matrix A_{2t} will have a dominant root λ_2/q_2 and we will write

$$q_2 = 1 + \alpha_2 \frac{N_2}{N_1}, \tag{6·2}$$

where $\alpha_2 > 0$ is another constant and N_1 the number of the species S_1 at time t. This equation expresses in a simple fashion the main biological consequences to the species S_2 of its dependence upon S_1 as a source of food. For when $N_1 \to 0$, $q_2 \to \infty$, and the inherent rate of increase of the predator $r_2' = \log_e(\lambda_2/q_2) \to -\infty$ (disappearance of predator in the absence of any prey). Conversely, when N_1 becomes very large, $q_2 \to 1$ and the inherent rate of increase of the predator approaches its optimum value $r_2 = \log_e \lambda_2$.

Adopting, then, the simple system represented by (6·1) and (6·2) we shall have for the stationary state, putting $q_1 = \lambda_1$ and $q_2 = \lambda_2$,

$$N_1 = \frac{\alpha_2(\lambda_1 - 1)}{\alpha_1(\lambda_2 - 1)} = K_1, \quad N_2 = \frac{\lambda_1 - 1}{\alpha_1} = K_2,$$

which will be real positive quantities when both λ_1 and $\lambda_2 > 1$. Moreover, assuming for the moment that a stable stationary state is possible, we must have $\alpha_2(\lambda_1 - 1) > \alpha_1(\lambda_2 - 1)$ and

$(\lambda_1-1)>\alpha_1$ for both species to coexist in appreciable numbers. Then, expressing α_1 and α_2 in terms of the λ's and K's,

$$q_1 = 1+(\lambda_1-1)\frac{N_2}{K_2}, \tag{6·1a}$$

$$q_2 = 1+(\lambda_2-1)\frac{K_1N_2}{K_2N_1}. \tag{6·2a}$$

This simple system, however, can be improved upon to some extent. It will be noticed that if in equation (6·1) $N_2 = 0$, $q_1 = 1$ and thus in the absence of the predator it is assumed that the prey will increase to an unlimited extent. In order to introduce the conception of a limited environment, we might put

$$q_1 = 1+\alpha_1N_2+\beta_1N_1, \tag{6·3}$$

so that when $N_2 = 0$, the species S_1 will approach some upper limit in numbers. A slightly more general system is represented then by equations (6·3) and (6·2), for which the stationary state is

$$N_1 = \frac{\alpha_2(\lambda_1-1)}{\alpha_1(\lambda_2-1)+\alpha_2\beta_1}, \quad N_2 = \frac{(\lambda_1-1)(\lambda_2-1)}{\alpha_1(\lambda_2-1)+\alpha_2\beta_1}.$$

It would thus be possible to examine the consequences of various hypotheses as to the way in which the reduction in the optimum inherent rates of increase for the two species are effected. The possible combinations are, however, so numerous that it is difficult to cover at all adequately any more than one of the most obvious cases. In order to illustrate the properties of such a system, the simplest, and also the possibly not unrealistic example of the reduction in the rates for both species taking place through the operation of an additional force of mortality independent of age will be considered here. That is to say, it will be assumed that the effect of the species S_2 on system of rates for the species S_1 will be to divide the elements of the matrix A_1 by the factor q_1, and similarly that the effect of the species S_1 on the species S_2 and the matrix A_2 will be to divide the elements of the latter by q_2. This simplifies a number of the actual computations and also the analysis of the properties of the equations.

If at time t the age distributions of the $N_1(t)$ and $N_2(t)$ individuals of the species S_1 and S_2 are of the stable forms appropriate to the dominant latent roots λ_1 and λ_2 of the matrices A_1 and A_2 respectively, then from the properties of a matrix $q^{-1}A$ which were discussed in the previous section, the two populations will retain their initial forms of age distribution unchanged. The total numbers of individuals in the two populations, supposing these are subject to the system defined by equations (6·1a) and (6·2a) respectively, will therefore be at time $t+1$

$$N_1(t+1) = \frac{\lambda_1N_1(t)}{1+(\lambda_1-1)\dfrac{N_2(t)}{K_2}},$$

$$N_2(t+1) = \frac{\lambda_2N_2(t)}{1+(\lambda_2-1)\dfrac{K_1N_2(t)}{K_2N_1(t)}},$$

whence

$$N_1(t+1)-N_1(t) = \frac{(\lambda_1-1)N_1(t)\left\{1-\dfrac{N_2(t)}{K_2}\right\}}{1+(\lambda_1-1)\dfrac{N_2(t)}{K_2}},$$

and

$$N_2(t+1)-N_2(t) = \frac{(\lambda_2-1)N_2(t)\left\{1-\dfrac{K_1N_2(t)}{K_2N_1(t)}\right\}}{1+(\lambda_2-1)\dfrac{K_1N_2(t)}{K_2N_1(t)}}.$$

Before discussing the limits to which these difference equations will tend when the time interval is made smaller and smaller, it is necessary to consider the question of the value to which the dominant latent root of the matrix will tend when the latter becomes of a very large order. Suppose that working in some convenient unit of age and time we have the matrix A_1, with a real positive root λ_1, representing some given system of age-specific fertility and mortality rates. We can also construct a new matrix—$A_{\frac{1}{2}}$ say—for the same system of rates when the time interval is taken to be a half-unit. This new matrix will be twice the order of the original one and it will have a dominant root—$\lambda_{\frac{1}{2}}$ say—which will be less than λ_1. Continuing the process further, we shall have for an interval of age and time h a matrix A_h with a dominant root λ_h, this root representing in the case of a population with a stable age distribution, the ratio $N(t+h)/N(t)$. In order to compare the successive values of λ_h which would be obtained by making the interval h smaller and smaller, it is necessary to express them in some common unit of time and we can write

$$\Lambda = (\lambda_h)^{1/h} \quad \text{or} \quad \lambda_h = \Lambda^h.$$

Then, when the matrix remains constant in time, we shall have for a population with a stable age distribution,

$$\frac{N(t+h)-N(t)}{h} = \frac{\lambda_h - 1}{h} N(t) = \frac{\Lambda^h - 1}{h} N(t),$$

or, when $h \to 0$,

$$\frac{dN}{dt} = (\log_e \Lambda)\, N,$$

since

$$\lim_{h \to 0} \frac{\Lambda^h - 1}{h} = \log_e \Lambda.$$

Thus, as the matrix is made larger and larger, the value of $\log_e \Lambda$ tends to ρ, the true instantaneous relative rate of increase of the stable population per unit of time.

In a similar fashion we may write for an interval h the above difference equations in the form

$$\frac{N_1(t+h)-N_1(t)}{h} = \frac{\left(\frac{\Lambda_1^h - 1}{h}\right) N_1(t) \left\{1 - \frac{N_2(t)}{K_2}\right\}}{1 + (\Lambda_1^h - 1)\frac{N_2(t)}{K_2}},$$

$$\frac{N_2(t+h)-N_2(t)}{h} = \frac{\left(\frac{\Lambda_2^h - 1}{h}\right) N_2(t) \left\{1 - \frac{K_1 N_2(t)}{K_2 N_1(t)}\right\}}{1 + (\Lambda_2^h - 1)\frac{K_1 N_2(t)}{K_2 N_1(t)}},$$

which, as $h \to 0$, may be replaced by

$$\frac{dN_1}{dt} = (\log_e \Lambda_1) N_1 \left(1 - \frac{N_2}{K_2}\right), \quad \frac{dN_2}{dt} = (\log_e \Lambda_2) N_2 \left(1 - \frac{K_1 N_2}{K_2 N_1}\right).$$

Thus, when the age distributions of the populations S_1 and S_2 are each initially of the appropriate stable form, and when it is assumed that their respective systems of rates are represented by the matrices $q_1^{-1}A_1$ and $q_2^{-1}A_2$, the system of interrelations between the two populations which is defined by

$$q_1 = 1 + \alpha_1 N_2, \quad q_2 = 1 + \alpha_2 \frac{N_2}{N_1},$$

is equivalent to that defined by the differential equations

$$\frac{dN_1}{dt} = (r_1 - a_1 N_2) N_1, \quad \frac{dN_2}{dt} = \left(r_2 - a_2 \frac{N_2}{N_1}\right) N_2, \tag{6·4}$$

or, when q_1 is defined by (6·3), to

$$\frac{dN_1}{dt} = (r_1 - a_1 N_2 - b_1 N_1) N_1, \quad \frac{dN_2}{dt} = \left(r_2 - a_2 \frac{N_2}{N_1}\right) N_2, \tag{6·5}$$

where in both sets $r_1 = \log_e \lambda_1$, $r_2 = \log_e \lambda_2$ and a_1, a_2, b_1 are constants > 0. This result is analogous to that discussed in the previous section for a single population increasing in a limited environment, where it was shown that when mortality was affected by a factor independent of age and the initial distribution was of the appropriate stable form, the numbers of individuals increased according to the logistic law, and that consequently under these conditions the type of population growth resulting from the operation of the matrix $q^{-1}A$, where

$$q = 1 + \frac{\lambda_1 - 1}{K} N,$$

was equivalent to that defined by the differential equation,

$$\frac{dN}{dt} = (r - aN) N.$$

The system of equations (6·4) differs somewhat from the classical Lotka-Volterra equations (Lotka, 1925, Chap. 8; Volterra, 1931, p. 14) for a simple predator-prey relationship between two species, in which the second member would be written

$$\frac{dN_2}{dt} = (-r_2 + a_2 N_1) N_2.$$

The form of the second member in (6·4) was originally suggested by the results of an analysis made by the author (unpublished observations) of some data given by Gause (1934) for the growth in numbers of *Paramecium caudatum* and *Paramecium aurelia* cultures, in which the food supply consisted of a suspension of *Bacillus pyocyaneus* in a buffered medium. Two different concentrations of bacteria—called by Gause 'one loop' and 'half-loop'—were used for both species of *Paramecium*, and under the conditions of the experiments these populations could be regarded as living in a limited environment with a constant supply of food. It was apparent from the results that for each species living alone the upper limit to the number of individuals depended on the concentration of food, being in each case approximately twice as great in the cultures with the 'one loop' concentration as in those with the 'half-loop'. If logistic equations are fitted to the four series of data given by Gause (1934, table 4, p. 145), it will be found that whereas the constant r in the equation $dN/dt = (r - aN) N$ remains approximately the same in the pair of experiments on each species of *Paramecium*, the constant a is inversely proportional to the concentration of food (see also on this point Kostitzin, 1937, p. 77). Thus, when the food supply (F) was kept constant, the form of population growth in numbers could be written

$$\frac{dF}{dt} = 0, \quad \frac{dN_2}{dt} = \left(r_2 - \frac{a_2}{C} N_2\right) N_2,$$

where C represents the relative concentration of food in the different experiments. This relationship suggested a system of equations such as (6·4) for the theoretical case of a food supply consisting of a population of individuals which when living alone would increase at

a rate $dN_1/dt = r_1 N_1$. However, apart from these considerations, the form of the second member of (6·4) is linked with the type of expression used here to define q_2 in terms of N_1 and N_2, and the latter arose as one of the simplest and most obvious ways of expressing the dependence of the species S_2 on S_1, bearing in mind that the elements of the matrix representing the system of rates at a given time must be positive quantities ($F_x \geqslant 0$, $0 < P_x \leqslant 1$). The difficulties which arise when this is not the case will be appreciated on endeavouring to find a working model in terms of matrices and vectors which will reduce to the classic Lotka-Volterra equations under suitable initial conditions. For, in the case of the predatory species S_2 we should have to consider a reciprocal matrix A_2^{-1} with a real positive root λ_2^{-1}, and at time t the matrix $q_2 A_2^{-1}$ would be regarded as operating on the vector $\xi(t)$ representing the age distribution of S_2. Then, if as before the matrix $q_1^{-1} A_1$ represents the system of rates for the species S_1 and

$$q_1 = 1 + \frac{\lambda_1 - 1}{K_2} N_2, \quad q_2 = 1 + \frac{\lambda_2 - 1}{K_1} N_1,$$

we have a system which will reduce to the Lotka-Volterra differential equations when the initial age distributions of both populations are of the stable form appropriate to their respective matrices A_1 and A_2. Now, apart from the fact that here no upper limit is placed on the inherent rate of increase, $r_2' = \log_e (q_2/\lambda_2)$, of the species S_2, there is an added complication that a number of the elements of A_2^{-1} will be negative (for the form of the matrix A^{-1} see the previous paper, § 4). Although no difficulties arise in the special case, when the age distribution of S_2 is of the stable form, in the perfectly general case of an arbitrary $\xi(t)$ some of the elements of $\xi(t+1) = q_2 A_2^{-1} \xi(t)$ can become negative and thus meaningless from the biological point of view. For these various reasons, therefore, the form of interrelationship between the two species defined by equations (6·1 *a*) and (6·2 *a*) was adopted here as a working model, and these reduce in the special case to the system of differential equations (6·4).

The writer has to confess that he has been unable to integrate either of the sets (6·4) and (6·5). Their main properties, however, seem to be quite clear. Taking the simplest system (6·4) first, we have for $dN_1/dt = dN_2/dt = 0$,

$$N_1 = \frac{r_1 a_2}{r_2 a_1} = K_1, \quad N_2 = \frac{r_1}{a_1} = K_2,$$

and, introducing for simplicity the variables $n_1 = N_1/K_1$, $n_2 = N_2/K_2$,

$$\frac{dn_1}{dt} = r_1 n_1 (1 - n_2), \quad \frac{dn_2}{dt} = r_2 n_2 \left(1 - \frac{n_2}{n_1}\right).$$

We will suppose that we are dealing with the case when $r_1 a_2 > r_2 a_1$, $r_1 > a_1$, in order that the stationary state may have a real meaning from the biological point of view. Then, in considering small departures from the stationary state, let $v_1 = n_1 - 1$, $v_2 = n_2 - 1$; and, disregarding in the usual way terms such as $v_1 v_2$, v_1^2, etc.;

$$\frac{dv_1}{dt} = -r_1 v_2, \quad \frac{dv_2}{dt} = r_2 v_1 - r_2 v_2.$$

This linear system will have a solution of the type $v_1 = A_1 e^{\mu_1 t} + B_1 e^{\mu_2 t}$, $v_2 = A_2 e^{\mu_1 t} + B_2 e^{\mu_2 t}$, where the values of μ will be given by the roots of the characteristic determinant

$$\begin{vmatrix} -\mu & -r_1 \\ r_2 & -(r_2 + \mu) \end{vmatrix} = 0$$

or

$$2\mu = -r_2 \pm \sqrt{(r_2^2 - 4 r_1 r_2)}.$$

Thus, both roots μ_1 and μ_2 will be complex so long as $r_2 < 4r_1$, and the real part of this pair will be negative since $r_2 > 0$. The system under these conditions will therefore approach the stationary state by a series of damped oscillations. When $r_2 > 4r_1$, both μ_1 and μ_2 will be negative; and consequently the stationary state will be stable, since in both cases v_1 and v_2 tend to zero as time increases.

The analysis of the system represented by equations (6·5) leads to very similar results. For

$$\frac{dN_1}{dt} = \frac{dN_2}{dt} = 0,$$

$$N_1 = \frac{r_1 a_2}{r_2 a_1 + a_2 b_1} = K_1, \quad N_2 = \frac{r_1 r_2}{r_2 a_1 + a_2 b_1} = K_2.$$

And, in the same way as before, putting $n_1 = N_1/K_1$, $n_2 = N_2/K_2$, $v_1 = n_1 - 1$, $v_2 = n_2 - 1$, and neglecting terms in $v_1 v_2$, etc., we have

$$\frac{dv_1}{dt} = -r_1(1-k)v_1 - r_1 k v_2, \quad \frac{dv_2}{dt} = r_2 v_1 - r_2 v_2,$$

where

$$k = \left(1 + \frac{a_2 b_1}{r_2 a_1}\right)^{-1} \quad (0 < k < 1).$$

Then, putting the characteristic determinant

$$\begin{vmatrix} -\{r_1(1-k)+\mu\} & -r_1 k \\ r_2 & -(r_2+\mu) \end{vmatrix} = 0,$$

we have

$$\mu^2 + \{r_2 + r_1(1-k)\}\mu + r_1 r_2 k = 0.$$

The roots of this equation will be either both negative or both complex with the real part negative, depending on the relation between the various constants, and consequently both v_1 and v_2 will tend to zero as time goes on, the stationary state thus being stable as before. It will be noticed, however, that if $\mu = u \pm iv$, the damping term represented by the real part, $u = -\{r_2 + r_1(1-k)\}$, will be greater than in the case of the first system of equations (6·4) where $u = -r_2$. Again, for a given set of values of r_1, r_2, a_1, a_2, the number of individuals $N_1 = K_1$, $N_2 = K_2$ must be less for the second set of equations than for the first, since by definition $b_1 > 0$. Thus we might expect that for a population subject to equations (6·5) the stationary numbers will be lower and the approach to the stationary state more rapid than for a population subject to equations (6·4), provided that the values of r_1, r_2, a_1 and a_2 are the same in both cases.

As a numerical example of these predator-prey equations, suppose that the optimum system of rates for two imaginary species were the same and that they were represented by the matrix A which has been used previously to illustrate various points. Then $A_1 = A_2$, $\lambda_1 = \lambda_2 = 3$ and $r_1 = r_2 = 1{\cdot}09861$. If, for the first set of equations (6·1) and (6·2) we put

$$q_1 = 1 + 0{\cdot}002 N_2, \tag{6·6}$$

$$q_2 = 1 + 10 N_2/N_1 \tag{6·7}$$

and for the second, (6·3) and (6·2)

$$q_1 = 1 + 0{\cdot}002 N_2 + 0{\cdot}000185185 N_1, \tag{6·8}$$

q_2 remaining as before, the number of individuals for the stationary state are in the first case $K_1 = 5000$, $K_2 = 1000$, and in the second $K_1 = 3418$, $K_2 = 684$. Then, assuming that at $t = 0$, $N_1 = N_2 = 108$, and that each of these populations had the same stable form of age distribution

$$\xi_1 = \{81, 21, 5, 1\},$$

the results of operating on these two age distributions with the matrices $q_1^{-1}A_1$ and $q_2^{-1}A_2$ were as follows for the two sets of equations. The first two columns give the numbers of prey (N_1) and predators (N_2) when no upper limit is placed on the number of prey (equations (6·6) and (6·7)), and the second two columns give the respective numbers when the upper limit to N_1 would be 10800 individuals, if the predatory species was absent (equations (6·8) and (6·7)). (This is the same logistic population as was used in the previous section as an illustration.)

	I		II	
t	N_1	N_2	N_1	N_2
0	108	108	108	108
1	266	30	262	30
2	755	42	710	42
3	2089	81	1754	79
4	5393	175	3550	163
5	11983	396	5371	335
6	20050	894	6046	619
7	21583	1854	5403	917
8	13756	2991	4227	1020
9	5910	2826	3317	897
10	2665	1466	2921	726
11	2033	677	2928	625
12	2592	469	3146	598
13	4012	501	3396	618
14	6011	668	3555	658
15	7717	949	3587	692
16	7986	1277	3529	709
17	6741	1474		
18	5122	1388		
19	4071	1122		
20	3763	896		
21	4043	795		
22	4684	804		

In both cases the approach to the stationary state by means of a series of damped oscillations is very evident, this approach being made more rapidly in the second series than in the first as was to be expected from the results of the foregoing analysis. Probably the clearest graphical illustration of these functions is obtained by plotting $\log N_2$ against $\log N_1$, the result being a spiral curve which gradually approaches the stationary point.

Although these predator-prey equations have been studied here only in the special case of the reduction in the rates of increase of the two populations being effected by an increase in the degree of mortality which is independent of age, there would be little difficulty in investigating, for instance, the type of case in which a relative absence of prey affected the fertility of the predator, and so forth. Moreover, there will be in all cases the effect on such a system of any abnormalities in the initial age distributions, or of any chance disturbances of the existing age distributions at some point in the development of the populations. Without working out any actual examples, however, it might be expected from the results obtained in the case of a logistic-type population that the general effect of all these factors would be to add further oscillatory features to those which already are inherent in the system itself, even when the stability of the age-distributions is established as in the above numerical examples. It seems likely, too, that these additional factors will increase the chance of one or other of the two species being reduced to such low numbers as would be equivalent in practice to the extinction of the population. This possibility will, however, greatly depend on the numerical relations between the various constants which enter into the equations

and upon the initial conditions of the particular system. Finally, just as in the case of a solitary population increasing in a limited environment, there is the possibility of studying the more complicated cases in which q_1 and q_2 are taken to be functions not only of N_1 and N_2 at time t, but of the numbers at some previous time, or of an integral of N_1 or N_2 between some time limits. Similar methods could also be used in order to study a chain of such predator-prey relations.

This work arose out of some research carried out by the Bureau of Animal Population with the aid of a grant from the Agricultural Research Council, to which body grateful acknowledgement is made.

REFERENCES

CHAPMAN, R. N. (1928). The quantitative analysis of environmental factors. *Ecology*, **9**, 111–22.
CROMBIE, A. C. (1945). On competition between different species of graminivorous insects. *Proc. Roy. Soc.* B, **132**, 362–95.
FISHER, R. A. (1930). *The Genetical Theory of Natural Selection*. Oxford: Clarendon Press.
GAUSE, G. F. (1934). *The Struggle for Existence*. Baltimore: Williams and Wilkins.
KOSTITZIN, V. A. (1937). *Biologie mathématique*. Paris: Armand Colin.
LESLIE, P. H. (1945). On the use of matrices in certain population mathematics. *Biometrika*, **33**, 183–212.
LEWIS, E. G. (1942). On the generation and growth of a population. *Sankhya*, **6**, 93–6.
LOTKA, A. J. (1925). *Elements of Physical Biology*. Baltimore: Williams and Wilkins.
LOTKA, A. J. (1931). The structure of a growing population. *Hum. Biol.* **3**, 459–93.
LOTKA, A. J. (1939). Théorie analytique des associations biologiques. II. Analyse démographique avec application particulière à l'espèce humaine. *Actualités Sci.* no. 780, 1–149. Paris: Hermann.
VOLTERRA, V. (1931). *Leçons sur la théorie mathématique de la lutte pour la vie*. Paris: Gauthier-Villars.
VOLTERRA, V. & D'ANCONA, U. (1935). Les associations biologiques au point de vue mathématique. *Actualités Sci.* no. 243, 1–96. Paris: Hermann.

THE TRANSFORMATION OF POISSON, BINOMIAL AND NEGATIVE-BINOMIAL DATA

By F. J. ANSCOMBE, *Rothamsted Experimental Station*

1. Introduction

Bartlett (1936) showed that if **r** is a Poisson variable with mean m and **y** is a random variable whose values y are derived by the transformation

$$y = \sqrt{r} \tag{1·1}$$

from the values r of **r**, then **y** is distributed rather more nearly normally than **r** with variance approximately $\frac{1}{4}$ if m is large, and the technique of analysis of variance may be applied to **y**.* He also showed that

$$y = \sqrt{(r+\tfrac{1}{2})} \tag{1·2}$$

is a better transformation, if slightly less convenient to use, as **y** then has more nearly a constant variance of $\frac{1}{4}$, even when m is not large. Similar transformations were proposed for a binomial variable, and Beall (1942) gave the transformation analogous to (1·1) appropriate to a negative binomial variable.

I begin by considering the transformation

$$y = \sqrt{(r+c)} \tag{1·3}$$

of a Poisson variable **r**, and show that for large m **y** has a most nearly constant variance (namely, $\frac{1}{4}$) when $c = \frac{3}{8}$; a result due to A. H. L. Johnson.

The similar transformation for a binomial variable **r**, with mean m and total number n, is

$$y = \sin^{-1}\sqrt{\left(\frac{r+c}{n+2c}\right)}. \tag{1·4}$$

The optimum value of c is $\frac{3}{8}$ if m and $n-m$ are large. The variance is approximately $\frac{1}{4}(n+\frac{1}{2})^{-1}$.

For a negative binomial variable **r**, with mean m and exponent k, the latter being constant and known, the corresponding transformation is

$$y = \sinh^{-1}\sqrt{\left(\frac{r+c}{k-2c}\right)}. \tag{1·5}$$

The optimum value of c is roughly $\frac{3}{8}$ if m is large and $k > 2$, and the variance is approximately $\frac{1}{4}\psi'(k)$, where $\psi'(t)$ denotes the second derivative of $\ln\Gamma(t)$ with respect to t. A simpler transformation, known to have an optimum property (i.e. to be the best of that degree of complexity) for m large and $k \geqslant 1$, is

$$y = \ln(r+\tfrac{1}{2}k); \tag{1·6}$$

the variance is approximately $\psi'(k)$. This is equivalent to setting $c = \frac{1}{2}k$ in (1·5). If k is large, $\psi'(k) = 1/(k-\frac{1}{2})$ approximately.

The effect of these transformations for small values of m is shown numerically. The value $\frac{3}{8}$ for c appears to be nearly optimum for practical purposes, except with the negative binomial distribution when k is small. (When $k = 2$, the optimum value of c appears to be about 0·2.) In any case, it may be more convenient to choose a one-decimal value for c, i.e. generally 0·3 or 0·4.

* Letters in heavy type denote random variables, of which the same letter in light type denotes a possible value. Some equations, in particular all those of §1, are equally valid for y and r in either type.

For making transformation (1·4), the table of the 'angular transformation' given by Fisher & Yates (1938) may be used, calculating the observed percentage as $100(r+c)/(n+2c)$, and replacing n by $n+\frac{1}{2}$ in the formula for the variance. For transformation (1·5), no corresponding table of $\sinh^{-1}\sqrt{x}$ has yet been published, I believe*.

2. POISSON DISTRIBUTION

We suppose **r** is distributed in a Poisson distribution with mean m, and we consider the transformation (1·3) above, with c a non-negative constant.

Let $r-m=t$ and $m+c=m'$. Coefficients a_s are defined for $s=1,2,3,\ldots$ by

$$a_s=(-1)^{s+1}\frac{1.(-1).(-3).\ldots.(-2s+3)}{2^s.s!}. \tag{2·1}$$

Then for any $t\geqslant -m'$ we have the Taylor series expansion

$$y=\sqrt{m'}\left\{1+a_1\frac{t}{m'}-a_2\left(\frac{t}{m'}\right)^2+\ldots+(-1)^s a_{s-1}\left(\frac{t}{m'}\right)^{s-1}\right\}+R_s. \tag{2·2}$$

If $t>0$, we see at once from Lagrange's form of the remainder term that R_s satisfies

$$|R_s|<\frac{a_s t^s}{(m')^{s-\frac{1}{2}}}. \tag{2·3}$$

Considering now $|t|\leqslant m'$, we have directly from (2·2)

$$R_s(m')^{-\frac{1}{2}}=\left(1+\frac{t}{m'}\right)^{\frac{1}{2}}-\left\{1+a_1\frac{t}{m'}-\ldots+(-1)^s a_{s-1}\left(\frac{t}{m'}\right)^{s-1}\right\}=\sum_{i=s}^{\infty}(-1)^{i+1}a_i\left(\frac{t}{m'}\right)^i. \tag{2·4}$$

The series converges, and we may write

$$\frac{R_s(m')^{s-\frac{1}{2}}}{t^s}=\sum_{i=s}^{\infty}(-1)^{i+1}a_i\left(\frac{t}{m'}\right)^{i-s}, \tag{2·5}$$

where the right-hand side again converges and is bounded. If $G(s)$ is a bound to its absolute magnitude,

$$|R_s|\leqslant G(s)\frac{|t|^s}{(m')^{s-\frac{1}{2}}}. \tag{2·6}$$

Comparing this inequality with (2·3), we see that it holds for all $t>-m'$.

We note now that the moments of **t** are

$$\mu_1=0,\quad \mu_2=m,\quad \mu_3=m,\quad \mu_4=3m^2+m,\quad \text{etc.}, \tag{2·7}$$

and the absolute moment of order n is $O(m^{\frac{1}{2}n})$ as $m\to\infty$. We may therefore take expectations formally in the right-hand side of (2·2) and its powers, and derive asymptotic expansions for the moments of **y** as $m\to\infty$. We find

$$\text{var}\,(\mathbf{y})\sim\frac{1}{4}\left\{1+\frac{3-8c}{8m}+\frac{32c^2-52c+17}{32m^2}\right\}, \tag{2·8}$$

so that when $c=\frac{3}{8}$

$$\text{var}\,(\mathbf{y})\sim\frac{1}{4}\left\{1+\frac{1}{16m^2}\right\}. \tag{2·9}$$

We have also

$$\text{E}\,(\mathbf{y})\sim\sqrt{(m+c)}-\frac{1}{8m^{\frac{1}{2}}}+\frac{24c-7}{128m^{\frac{3}{2}}}. \tag{2·10}$$

If we set

$$\text{E}\,(\mathbf{y})=\sqrt{(m_y+c)}, \tag{2·11}$$

* [Beall (1942, pp. 250–51) gave a table of $x^1=k^{-\frac{1}{2}}\sinh^{-1}(kx)^{\frac{1}{2}}$, suited for the form of transformation which he used. ED.]

m_y is the estimate of m derived by applying the transformation (1·3) in reverse to the arithmetic mean $\bar{y}$ of a large sample of observed values of **y**. From (2·10) we have

$$m_y \sim m - \frac{1}{4} + \frac{8c-3}{32m}, \tag{2·12}$$

so that setting $c = \frac{3}{8}$ also renders the bias $m_y - m$ in m_y nearly constant.

Skewness and kurtosis of the distribution of **y** are measured by

$$\gamma_1 \sim -\frac{1}{2m^{\frac{1}{2}}}\left\{1 + \frac{25-48c}{16m}\right\}, \tag{2·13}$$

$$\gamma_2 \sim \frac{1}{m}\left\{1 + \frac{945-1536c}{256m}\right\}. \tag{2·14}$$

These compare with $m^{-\frac{1}{2}}$ and m^{-1} for the original Poisson variable **r**.

It is also of interest to find the large-sample efficiency E_y of the arithmetic mean $\bar{y}$ of observed values of **y** as a statistic for determining m (Bartlett, 1936, 1947). If **x** is a random variable having a distribution (absolutely continuous or discrete) such that the arithmetic mean $\bar{x}$ is a sufficient statistic for determining a parameter θ, it is easy to show, using the form of the frequency function of **x** given by Fisher (1934), that the large-sample efficiency E_y of the average $\bar{y}$ of any function **y** of **x**, for determining θ, is the square of the correlation coefficient between x and y, i.e.

$$E_y = \frac{[\mathrm{cov}\,(\mathbf{x},\mathbf{y})]^2}{\mathrm{var}\,(\mathbf{x}).\,\mathrm{var}\,(\mathbf{y})}. \tag{2·15}$$

In the present instance,
$$E_y = \frac{[\mathrm{cov}\,(\mathbf{r},\mathbf{y})]^2}{m\,\mathrm{var}\,(\mathbf{y})} \sim 1 - \frac{1}{8m} + \frac{16c-9}{64m^2}. \tag{2·16}$$

3. Binomial distribution

We suppose now that **r** is distributed in a binomial distribution with total number n and mean m $(0 < m < n)$, and we consider the transformation

$$y = \sqrt{(n+d_2)}\sin^{-1}\sqrt{\left(\frac{r+c}{n+d_1}\right)}, \tag{3·1}$$

where c, d_1, d_2 are constants to be determined. Setting

$$r - m = t, \quad m + c = m', \quad n + d_1 = n_1, \quad n + d_2 = n_2,$$

the transformation becomes
$$y = \sqrt{n_2}\sin^{-1}\sqrt{\left(\frac{m'+t}{n_1}\right)}. \tag{3·2}$$

We can expand y in a Taylor series in ascending powers of t, for $-m' \leqslant t \leqslant n_1 - m'$, and show that R_s, the remainder after s terms, is such that

$$\frac{n_1^s R_s}{t^s \sqrt{n_2}}$$

is bounded for all t in the range considered, with bounds depending on s and the ratio m'/n_1 only. Thus

$$|R_s| \leqslant G\left(s, \frac{m'}{n_1}\right)\frac{|t|^s}{n_1^s}\sqrt{n_2}. \tag{3·3}$$

Hence we can obtain asymptotic expansions of the moments of y, valid for large n and constant ratio m/n. We find

$$\text{var}(\mathrm{y}) \sim \frac{1}{4}\left\{1+\frac{2d_2-1}{2n}+\frac{3-8c}{8m}+\frac{3+8c-8d_1}{8(n-m)}\right\}, \tag{3·4}$$

so that if we choose $c=\frac{3}{8}$, $d_1=\frac{3}{4}$, $d_2=\frac{1}{2}$, we have

$$\text{var}(\mathrm{y}) = \tfrac{1}{4}+O\left(\frac{1}{n^2}\right). \tag{3·5}$$

It will be noticed here that $d_1 = 2c$, so that the transformation is symmetrical about $r=\frac{1}{2}n$. The choice of d_2 only affects the scale of y (for n fixed), and not the constancy of var (y) as m varies, nor the shape of the distribution of y.

The equations corresponding to (2·12), (2·13), (2·14) and (2·16) are (proceeding to one term fewer in each case, for simplicity)

$$m_y \sim m+\frac{2m-n}{4n}, \tag{3·6}$$

$$\gamma_1 \sim \frac{2m-n}{2\{nm(n-m)\}^{\frac{1}{2}}}, \tag{3·7}$$

$$\gamma_2 \sim \frac{n^2-2m(n-m)}{nm(n-m)}, \tag{3·8}$$

$$E_y \sim 1-\frac{(2m-n)^2}{8nm(n-m)}. \tag{3·9}$$

4. NEGATIVE BINOMIAL DISTRIBUTION

We can deal similarly with a negative binomial variable r, with mean m and exponent k $(m, k>0)$, such that the probability of observing a value r is

$$p_r = \frac{\Gamma(r+k)}{r!\,\Gamma(k)}\left(\frac{m}{m+k}\right)^r\left(1+\frac{m}{k}\right)^{-k} \quad (r=0,1,2,\ldots). \tag{4·1}$$

The asymptotic expansions will be valid for large m and constant ratio k/m. We find the transformation

$$y = \sqrt{(k-\tfrac{1}{2})}\sinh^{-1}\sqrt{\left(\frac{r+\frac{3}{8}}{k-\frac{3}{4}}\right)}, \tag{4·2}$$

with

$$\text{var}(\mathrm{y}) = \tfrac{1}{4}+O\left(\frac{1}{m^2}\right). \tag{4·3}$$

But it is of more interest to consider m large and k fixed. (The corresponding problem does not arise with the positive binomial, since $m<n$ necessarily.) The preceding method of obtaining the expansions now breaks down, as it relied on the ratio of standard deviation to mean of r tending to zero as $m\to\infty$. Now we have the ratio tending to $k^{-\frac{1}{2}}$.

We consider two transformations,

$$y = 2\sinh^{-1}\sqrt{\left(\frac{r+c}{k+d}\right)} \tag{4·4}$$

and

$$y = \ln(r+A). \tag{4·5}$$

It is supposed that c, $k+d$, and A are positive and constant. Apart from an added constant, (4·4) may be written

$$y = 2\ln\{\sqrt{(r+c)}+\sqrt{(r+c+k+d)}\}. \tag{4·6}$$

When r is large, we have (again ignoring an added constant)

$$y \sim \ln r + \frac{A}{r} - \frac{B^2}{2r^2}, \tag{4·7}$$

where for (4·4) or (4·6)

$$A = \tfrac{1}{2}(2c+k+d), \quad B^2 = \tfrac{1}{8}\{8c^2+8c(k+d)+3(k+d)^2\}, \tag{4·8}$$

and for (4·5)

$$B = A. \tag{4·9}$$

We proceed to find an asymptotic expansion (as $m \to \infty$, with k fixed) for the moment-generating function of y, i.e. for

$$M(t) \equiv \sum_{r=0}^{\infty} e^{vt} p_r. \tag{4·10}$$

We set

$$\frac{m}{m+k} = e^{-\alpha}, \tag{4·11}$$

so that $\alpha \to 0$ as $m \to \infty$; and

$$e^{vt} \frac{\Gamma(r+k)}{\Gamma(k)\, r!} e^{-r\alpha} = u_r(\alpha). \tag{4·12}$$

We first prove the following

LEMMA. *As* $\alpha \to 0$,

$$\sum_{r=0}^{\infty} u_r(\alpha) - \int_0^{\infty} u_r(\alpha)\, dr \tag{4·13}$$

tends to a finite limit (depending on k *and* t*, and on which function* y *of* r *is chosen, namely* (4·4) *or* (4·5)).

Proof. By differentiating $\ln u_r(\alpha)$ with respect to r we see that as $r \to \infty$

$$\left(\frac{\partial}{\partial r}\right)^p u_r(\alpha) = u_r(\alpha)\left\{\frac{C}{r^p} + O\left(\frac{1}{r^{p+1}}\right)\right\} \quad (p = 1, 2, \ldots), \tag{4·14}$$

where C depends on $t+k-1$ and p but not on r or α, and the expression $O(r^{-(p+1)})$ is valid uniformly as $\alpha \to 0$. Since for large r

$$u_r(\alpha) = \frac{r^{t+k-1} e^{-r\alpha}}{\Gamma(k)}\left(1 + O\left(\frac{1}{r}\right)\right), \tag{4·15}$$

we see that

$$\left(\frac{\partial}{\partial r}\right)^p u_r(\alpha) = C e^{-r\alpha} r^{t+k-1-p}\left(1 + O\left(\frac{1}{r}\right)\right) \tag{4·16}$$

as $r \to \infty$, uniformly as $\alpha \to 0$, where C satisfies the same condition as before. If $p > t+k+1$, we may write

$$\left|\left(\frac{\partial}{\partial r}\right)^p u_r(\alpha)\right| < \frac{C}{(r+1)^2}, \tag{4·17}$$

where C is independent of α, and this holds for all $r \geqslant 0$ because the left-hand side is bounded for r in any bounded interval.

For all p, $(\partial/\partial r)^p u_r(\alpha) \to 0$ as $r \to \infty$, while at $r = 0$ $(\partial/\partial r)^p u_r(\alpha)$ is finite and tends to a finite limit as $\alpha \to 0$. We therefore have the Euler-Maclaurin expansion

$$\sum_{r=0}^{\infty} u_r(\alpha) = \int_0^{\infty} u_r(\alpha)\, dr + \tfrac{1}{2} u_0(\alpha) - \sum_{\nu=1}^{s-1} \frac{B_{2\nu}}{(2\nu)!}\left(\frac{\partial}{\partial r}\right)^{2\nu-1} u_0(\alpha) + R_s, \tag{4·18}$$

where

$$R_s = \int_0^{\infty} \theta(r)\left(\frac{\partial}{\partial r}\right)^{2s} u_r(\alpha)\, dr, \tag{4·19}$$

and $\theta(r)$ is a bounded periodic function of r, depending on s only. k is given, and we may restrict t to a neighbourhood of zero, say $|t| < \epsilon$. We choose s to be an integer satisfying $2s > k+1+\epsilon$. Then from (4·17) we see that R_s tends to a finite limit as $\alpha \to 0$, as do all the other terms after the first on the right-hand side of (4·18). Hence the lemma is proved.

Applying the lemma to (4·10), we have, as $\alpha \to 0$,

$$M(t) = (1-e^{-\alpha})^k \int_0^\infty u_r(\alpha)\, dr + O(\alpha^k). \tag{4·20}$$

In general the integral cannot be evaluated exactly. We may, however, expand $u_r(\alpha)$ for large r asymptotically, the first term being shown in (4·15) above. The error of the expansion is always less than a multiple of the next term (independent of α) for $r \geqslant 1$. Integrating term by term between the limits 1 and ∞, we obtain after a tedious reduction the following

THEOREM. *With the definitions of* (4·1), (4·7), (4·10) *and* (4·11), $M(t)$ *can be expanded asymptotically for* $\alpha \to 0$ *in the form*

$$M(t) = \frac{\Gamma(k+t)}{\alpha^t \Gamma(k)} \Big[1 + (A - \tfrac{1}{2}k)\, t \frac{\alpha}{k+t-1} + \{(\tfrac{1}{2}(A-\tfrac{1}{2}k)^2 + \tfrac{1}{24}k)\, t^2 + (\tfrac{1}{2}kA - \tfrac{1}{24}k(k+3) - \tfrac{1}{2}B^2)\, t\} \frac{\alpha^2}{(k+t-1)(k+t-2)} + \ldots \Big] + O(\alpha^k). \tag{4·21}$$

The series in square brackets is continued as far as the term in α^n, *where* n *is the greatest integer less than* k. t *is supposed confined to a neighbourhood of zero.*

The cumulant-generating function is now found by taking the logarithm of $M(t)$, namely,

$$K(t) = -t \ln \alpha + \ln \Gamma(k+t) - \ln \Gamma(k) + (A - \tfrac{1}{2}k)\, t \frac{\alpha}{k+t-1} + \ldots. \tag{4·22}$$

Hence
$$\operatorname{var}(\mathbf{y}) \sim \psi'(k) + \frac{k-2A}{(k-1)^2}\alpha, \tag{4·23}$$

if $k > 1$. We use the notation $\psi(t)$, $\psi'(t)$, etc., for the successive derivatives of $\ln \Gamma(k)$. If $A = \frac{1}{2}k$, we have

$$\operatorname{var}(\mathbf{y}) \sim \psi'(k) + \frac{k(k-1)(k-2) - (2k-3)(5k^2 - 3k - 12B^2)}{12(k-1)^2(k-2)^2}\alpha^2, \tag{4·24}$$

if $k > 2$. Considering $\mathbf{y}$ defined by (4·4), the condition $A = \frac{1}{2}k$ gives $d = -2c$, and the coefficient of α^2 in $\operatorname{var}(\mathbf{y})$ vanishes if c takes a value dependent on k, which for large k is approximately

$$c = \frac{3}{8} + \frac{23}{192}\frac{1}{k}, \tag{4·25}$$

and which rises to a little above 0·4 as k decreases towards 2. It appears from the numerical work below that a value of c somewhat under $\frac{3}{8}$ is optimum for practical purposes (if $k \geqslant 2$). For $\mathbf{y}$ defined by (4·5), we set $A = \frac{1}{2}k$ and then

$$\operatorname{var}(\mathbf{y}) \sim \psi'(k) - \frac{k(3k^2 - 9k + 7)}{12(k-1)^2(k-2)^2}\alpha^2, \tag{4·26}$$

if $k > 2$. If k is large and $m \gg k$, we have $\alpha \sim k/m$ and

$$\operatorname{var}(\mathbf{y}) \sim \frac{1}{k}\left(1 - \frac{k^2}{4m^2}\right). \tag{4·27}$$

Thus the larger k is the larger m must be for $\operatorname{var}(\mathbf{y})$ to approach its limiting value when $m \to \infty$. The transformation (4·5) is therefore not satisfactory if k is large.

For either form of transformation, we find the following limiting values as $m \to \infty$ ($\alpha \to 0$):

$$m_y/m = \exp\{\psi(k) - \ln k\}, \tag{4·28}$$

$$\gamma_1 = \psi''(k)/[\psi'(k)]^{\frac{3}{2}}, \tag{4·29}$$

$$\gamma_2 = \psi'''(k)/[\psi'(k)]^2. \tag{4·30}$$

The effect of the transformation on the shape of the distribution is the same as that of the logarithmic transformation of a variance estimate of $2k$ degrees of freedom, as studied by Bartlett & Kendall (1946). We find also, by a similar process, the limiting value of the efficiency of $\bar{y}$ in estimating m, namely,

$$E_y = [k\psi'(k)]^{-1}. \tag{4·31}$$

For $\alpha > 0$, these expressions hold with error $O(\alpha^2)$ when $A = \frac{1}{2}k$, if $k > 2$.

To conclude this investigation of asymptotic properties, we consider the transformation (4·5) when $k = 1$. We note first that if $k < 2$, and t is suitably restricted, the limit as $\alpha \to 0$ of (4·13) is also the limit as $N \to \infty$ of

$$\sum_{r=0}^{N} u_r(0) - \int_0^{N+\frac{1}{2}} u_r(0)\,dr. \tag{4·32}$$

When $k = 1$,

$$u_r(0) = (r + A)^t, \tag{4·33}$$

and the limit is easily evaluated as

$$f(A, t) = \tfrac{1}{2} + \ln\{\sqrt{(2\pi)}\,A^A e^{-A}/\Gamma(A)\}\,t + O(t^2). \tag{4·34}$$

We also note that the derivative of (4·13) with respect to α is finite and tends to a finite limit as $\alpha \to 0$ (proof as for the lemma itself), so that the error in replacing (4·13) by its limit as $\alpha \to 0$ is $O(\alpha)$. We thus find

$$M(t) = f(A, t)\,\alpha + (1 - e^{-\alpha})\int_0^{\infty} (r + A)^t e^{-r\alpha}\,dr + O(\alpha^2). \tag{4·35}$$

The integral is an incomplete Γ-function and is easily expanded in powers of α. Hence

$$\begin{aligned}\operatorname{var}(\mathbf{y}) = {} & \psi'(1) - (A - \tfrac{1}{2})\,\alpha(\ln\alpha)^2 \\ & + 2\{(A - \tfrac{1}{2})\,\psi(1) + A - A\ln A + \ln(\sqrt{(2\pi)}\,A^A e^{-A}/\Gamma(A))\}\,\alpha\ln\alpha + O(\alpha).\end{aligned} \tag{4·36}$$

Setting $A = \frac{1}{2}$ (in agreement with our previous rule $A = \frac{1}{2}k$), and $\alpha = m^{-1}$, we have

$$\operatorname{var}(\mathbf{y}) = \psi'(1) - \ln 2\,\frac{\ln m}{m} + O\left(\frac{1}{m}\right). \tag{4·37}$$

5. Numerical investigation

Table 1 refers to the transformations (1·3), (1·4) and (1·5), with $c = \frac{3}{8}$ in each case. The first section gives the error of the estimate m_y of m got by applying the transformation in reverse to $\mathbf{E}(\mathbf{y})$. The second section gives the ratio of $\operatorname{var}(\mathbf{y})$ to the 'limiting variance'. The limiting variances are respectively $\frac{1}{4}$, $\frac{1}{4}(n + \frac{1}{2})^{-1}$ and $\frac{1}{4}\psi'(k)$; the first and third of these being the actual limits as $m \to \infty$, and the second the value suggested in § 3 for n large. The remaining sections of Table 1 give the shape constants γ_1 and γ_2 of the distribution of y and the efficiency E_y. With the binomial distribution, no entries are given for $m > \frac{1}{2}n$, as the functions are obviously symmetric or skew-symmetric about $m = \frac{1}{2}n$.

For the Poisson distribution, the variances may be compared with those given by Bartlett (1936) for $c = 0$ and $\frac{1}{2}$. For the negative binomial, Table 2 gives $\operatorname{var}(\mathbf{y})$ when $c = 0$ and $\frac{1}{2}$, and $k = 2$; and also $\operatorname{var}(\mathbf{y})$ for the transformation (1·6), when $k = 2$ and 5, the limiting variance being now $\psi'(k)$.

Comparing (1·5) and (1·6), we have indicated above that (1·5) is more effective in stabilizing the variance if $k > 2$, and the computations confirm this. If $c = \frac{3}{8}$, (1·5) is defined for all $k > \frac{3}{4}$, and it is quite possible that throughout this range it is superior to (1·6). As $k \to \frac{3}{4}$ (or, more generally, as $k \to 2c$) the two transformations become equivalent, and (1·6) is defined for all $k > 0$. Bearing in mind that (1·6) is more convenient to use than (1·5), it seems reasonable to recommend that (1·6) should be used if $k < 2$, and also for larger k if m is large; and otherwise (1·5).

Table 1. *Transformations with* $c = \frac{3}{8}$

m	Poisson (1·3)	Binomial (1·4)		Negative binomial (1·5)	
		$n=10$	$n=5$	$k=5$	$k=2$
		Bias in the mean $(m_y - m)$			
1	−0·179	−0·144	−0·107	−0·244	−0·330
2	−0·231	−0·142	−0·047	−0·395	−0·609
3	−0·246	−0·100	—	−0·516	−0·869
4	−0·250	−0·051	—	−0·625	−1·121
6	−0·251	—	—	−0·830	−1·612
10	−0·250	—	—	−1·227	−2·577
20	−0·250	—	—	−2·209	−4·960
∞	−0·250	—	—	$-0{\cdot}098m$	$-0{\cdot}237m$
		Variance as fraction of the limiting variance			
1	0·717	0·759	0·802	0·639	0·537
2	0·924	0·958	0·990	0·848	0·731
3	0·983	1·001	—	0·929	0·824
4	0·999	1·006	—	0·964	0·875
6	1·002	—	—	0·990	0·927
10	1·001	—	—	0·999	0·966
20	1·000	—	—	1·000	0·989
∞	1·000	—	—	1·000	1·000
		Skewness coefficient γ_1			
1	0·31	0·30	0·31	0·34	0·41
2	−0·10	−0·07	0·01	−0·10	−0·04
3	−0·23	−0·13	—	−0·28	−0·25
4	−0·25	−0·08	—	−0·37	−0·37
6	−0·22	—	—	−0·45	−0·51
10	−0·17	—	—	−0·48	−0·63
20	−0·11	—	—	−0·48	−0·72
∞	0·00	—	—	−0·47	−0·78
		Kurtosis coefficient γ_2			
1	−0·80	−0·70	−0·54	−0·92	−1·0∠
2	−0·40	−0·21	0·06	−0·67	−0·93
3	−0·02	0·16	—	−0·32	−0·68
4	0·15	0·25	—	−0·06	−0·44
6	0·20	—	—	0·24	−0·08
10	0·12	—	—	0·43	0·35
20	0·05	—	—	0·46	0·77
∞	0·00	—	—	0·44	1·19
		Efficiency E_y			
1	96·5	97·7	98·9	93·7	89·9
2	95·8	97·9	99·4	91·3	85·5
3	96·1	98·9	—	90·2	83·0
4	96·7	99·5	—	89·9	81·7
6	97·8	—	—	89·8	80·2
10	98·7	—	—	89·9	78·9
20	99·4	—	—	90·2	78·0
∞	100·0	—	—	90·4	77·5

Table 2. *Other transformations of the negative binomial distribution*

m	Transformation (1·5) $k=2$		Transformation (1·6)	
	$c=0$	$c=\frac{1}{2}$	$k=2$	$k=5$
	Variance as fraction of the limiting variance			
1	1·106	0·503	0·456	0·369
2	1·272	0·697	0·648	0·569
3	1·255	0·793	0·749	0·689
4	1·229	0·849	0·811	0·766
6	1·174	0·908	0·879	0·853
10	1·106	0·954	0·936	0·928
20	1·045	0·984	0·976	0·977
∞	1·000	1·000	1·000	1·000

The transformation (1·3), with $c = \frac{3}{8}$, derived by the formal expansion given in § 2, was communicated to me by Mr A. H. L. Johnson. He has kindly agreed to my publishing his result. I am indebted to Mr L. K. Turner and Dr A. A. Rayner for their patience in carrying out the laborious computations.

REFERENCES

Bartlett, M. S. (1936). The square root transformation in the analysis of variance. *J. R. Statist. Soc. Suppl.* **3**, 68.

Bartlett, M. S. (1947). The use of transformations. *Biometrics*, **3**, 39.

Bartlett, M. S. & Kendall, D. G. (1946). The statistical analysis of variance-heterogeneity and the logarithmic transformation. *J. R. Statist. Soc. Suppl.* **8**, 128.

Beall, G. (1942). The transformation of data from entomological field experiments so that the analysis of variance becomes applicable. *Biometrika*, **32**, 243.

Fisher, R. A. (1934). Two new properties of mathematical likelihood. *Proc. Roy. Soc.* A, **144**, 285.

Fisher, R. A. & Yates, F. (1938). *Statistical Tables for Biological, Agricultural and Medical Research* (3rd ed. 1948). Edinburgh: Oliver and Boyd.

SOME THEOREMS ON TIME SERIES

II. THE SIGNIFICANCE OF THE SERIAL CORRELATION COEFFICIENT*

By P. A. P. MORAN, *Institute of Statistics, Oxford University*

In this note we show how to calculate exact values of the lower moments of the sampling distributions of various definitions of the serial correlation coefficient under the assumption that successive elements of the time series are in fact independent.

Supposing that $x_1, \ldots, x_n$ are the successive values of a time series, we may define, following Orcutt (1948), the sample serial correlation coefficient r_s as

$$\frac{n}{n-s}\frac{\sum_1^{n-s}(x_i-\bar{x})(x_{i+s}-\bar{x})}{\sum_1^n (x_i-\bar{x})^2}, \tag{1}$$

where $n\bar{x} = \sum_1^n x_i$, and we assume the x's are independently distributed in the same distribution with finite moments. We write $z_i = x_i - \bar{x}$. Then the z's are distributed in a singular multivariate distribution such that the correlation between any two z's is $-(n-1)^{-1}$. The first method we shall use for finding the moments of the distribution of (1) depends on observing that $\sum_1^n z_i = 0$ and that the z_i are all on an equal footing. When there really is serial correlation in the time series this latter fact ceases to be true and the method breaks down. Consider first the expectation of r_s:

$$\begin{aligned} E(r_s) &= \frac{n}{n-s}E\left\{\frac{\sum_1^{n-s} z_i z_{i+s}}{\sum_1^n z_i^2}\right\} \\ &= nE\left\{\frac{z_1 z_{1+s}}{\sum_1^n z_i^2}\right\} \\ &= \frac{1}{n-1}E\left\{\frac{\sum_{i+j} z_i z_j}{\sum_1^n z_i^2}\right\} \\ &= \frac{1}{n-1}E\left\{\frac{\left(\sum_1^n z_1\right)^2 - \sum_1^n z_i^2}{\sum_1^n z_i^2}\right\} \\ &= -(n-1)^{-1}. \end{aligned} \tag{2}$$

R. L. Anderson (1942) and W. J. Dixon (1944) have discussed the distribution of the serial correlation coefficient using the cyclic definition suggested by H. Hotelling, namely,

$$r_{cs} = \frac{\sum_1^n z_i z_{i+s}}{\sum_1^n z_i^2}, \tag{3}$$

* Paper I in this series was published in *Biometrika*, **34**, 281–91, (1947).

where z_{n+t} is defined to be equal to z_t for $t = 1, 2, \ldots$ and we add a suffix c to distinguish this definition from the previous one. By arguing in the same manner as above we see that

$$E(r_{cs}) = -(n-1)^{-1}. \tag{4}$$

This is identical with the result given by W. J. Dixon (1944). We note that formulae (2) and (4) are independent of s and do not require the assumption that the values $x_1 \ldots x_n$ are distributed normally. From now on we assume normality and we confine our attention to the distributions of r_1 and r_{c1}. r_s and r_{cs} $(s > 1)$ could be discussed in a similar manner, but the assumption of normality is essential in what follows.

The second and third moments can be found by two different methods which have been used to check each other. Consider first the second moment. For samples from a normal population it is known (e.g. M. G. Kendall, 1943, p. 289) that the 'standardized' fourth k-statistic, $k_4 k_2^{-2}$ is distributed independently of k_2 and so $E(k_4 k_2^{-2})\, E(k_2^2) = E(k_4) = 0$, giving $E(k_4 k_2^{-2}) = 0$. We now use the notation

$$[p_1^{r_1} \ldots p_k^{r_k}] = \Sigma z_1^{p_1} z_2^{p_1} \ldots z_s^{p_k},$$

where the sum is taken over all unequal values of the suffixes $i, j, \ldots, k$, so that, for example,

$$[4] = \sum_1^n z_i^4, \quad [2^2] = \sum_{i \neq j} z_i^2 z_j^2.$$

Then
$$\frac{k_4}{k_2^2} = \frac{(n-1)}{(n-2)(n-3)}\left[n(n+1)\frac{[4]}{[2]^2} - 3(n-1)\right],$$

and so
$$E\frac{[4]}{[2]^2} = \frac{3(n-1)}{n(n+1)}.$$

We also have, since $[1] = 0$, such relations as

$$[2^2] = [2]^2 - [4], \quad [21^2] = 2[4] - [2]^2, \quad [1^4] = 3[2]^2 - 6[4].$$

Now write $I = \dfrac{n-1}{n} r_1$ and consider

$$E(I^2) = E\left\{\frac{\sum_1^{n-1} z_i z_{i+1}}{\sum_1^n z_i^2}\right\}^2$$

$$= E\left\{\sum_1^n z_i^2\right\}^{-2}\left\{\sum_1^{n-1} z_i^2 z_{i+1}^2 + 2\sum_1^{n-2} z_i z_{i+1}^2 z_{i+2} + \Sigma z_i z_{i+1} z_j z_{j+1}\right\},$$

where the third sum in the numerator is taken over all sets of suffixes i and j such that $i, i+1$, $j, j+1$ are distinct. Since the z's are all on an equal footing, the above expression is equal to

$$E\left\{\sum_1^n z_i^2\right\}^{-2}\{(n-1) z_1^2 z_2^2 + 2(n-2) z_1^2 z_2 z_3 + (n-2)(n-3) z_1 z_2 z_3 z_4\}$$

$$= E[2]^{-2}\left\{\frac{1}{n}[2^2] + \frac{2}{n(n-1)}[21^2] + \frac{1}{n(n-1)}[1^4]\right\}$$

$$= E\left\{\frac{1}{n}\left(1 - \frac{[4]}{[2^2]}\right) + \frac{2}{n(n-1)}\left(\frac{2[4]}{[2]^2} - 1\right) + \frac{1}{n(n-1)}\left(3 - \frac{6[4]}{[2]^2}\right)\right\}$$

$$= \frac{n^2 - 3n + 3}{n^2(n-1)}$$

Then
$$E\{I - E(I)\}^2 = \frac{(n-2)^2}{n^2(n-1)},$$

and so
$$\operatorname{var} r_1 = \frac{(n-2)^2}{(n-1)^3}.$$

If we had taken the cyclic definition we should have found

$$E(r_{c1}^2) = E\left(\sum_1^n z_i^2\right)^{-2}\{nz_1^2z_2^2+2nz_1^2z_2z_3+n(n-3)z_1z_2z_3z_4\}$$
$$= E\left(\frac{1}{n-1}\left(1-\frac{[4]}{[2]^2}\right)+\frac{2}{(n-1)(n-2)}\left(\frac{2[4]}{[2]^2}-1\right)+\frac{1}{(n-1)(n-2)}\left(3-\frac{6[4]}{[2]^2}\right)\right)$$
$$= (n+1)^{-1} \quad \text{agreeing with Dixon (1944),}$$

and so
$$\text{var}\,(r_{c1}) = \frac{n(n-3)}{(n+1)(n-1)^2}. \tag{5}$$

A similar method can be used to find var (r_s) and var (r_{cs}).

Now consider the third moment:

$$E(I^3) = E\left(\sum_1^n z_i^2\right)^{-3}\left\{\sum_1^{n-1} z_i^3z_{i+1}^3+3\sum_1^{n-2} z_i^2z_{i+1}^3z_{i+2}+3\sum_1^{n-2} z_iz_{i+1}^3z_{i+2}^2\right.$$
$$+6\sum_1^{n-3} z_iz_{i+1}^2z_{i+2}^2z_{i+3}+6\sum z_iz_{i+1}^2z_{i+2}z_jz_{j+1}$$
$$\left.+3\sum z_i^2z_{i+1}^2z_jz_{j+1}+\sum z_iz_{i+1}z_jz_{j+1}z_kz_{k+1}\right\}, \tag{6}$$

where the last three sums are taken over all values of i, j and k which make the suffixes distinct. Then

$$E(I^3) = E\left(\sum_1^n z_i^2\right)^{-3}\{(n-1)z_1^3z_2^3+6(n-2)z_1^3z_2^2z_3+6(n-3)z_1^2z_2^2z_3z_4$$
$$+6(n-3)(n-4)z_1^2z_2z_3z_4z_5+3(n-2)(n-3)z_1^2z_2^2z_3z_4+(n-3)(n-4)(n-5)z_1z_2z_3z_4z_5z_6\}.$$

Then to apply the previous method we could express the ratios of the individual terms in the numerator to the denominator in terms of $[6][2]^{-3}$, $[4][2]^{-2}$ and $[3]^2[2]^{-3}$ and evaluate the expectation of these by using k-statistics.* This, in fact, I did as a check, but it is easier to use another method of argument based on the fact, pointed out by Koopmans (1942), that r_1 (and r_{c1}) is distributed independently of $\sum_1^n z_i^2$. This follows when we observe that $\sum_1^n z_i^2$ is distributed as χ^2 with $n-1$ degrees of freedom. We could therefore transform the z's into variates t_i $(i = 1, \ldots, n-1)$ such that

$$\sum_1^n z_i^2 = \sum_1^{n-1} t_i^2,$$

where the t's are independently distributed in the same normal distribution with zero mean. Then I (and r_c) is transformed into the ratio of a quadratic form in the t's to $\sum_1^{n-1} t_i^2$ and therefore depends only on the angles which the vector $(t_1 \ldots t_{n-1})$ make with the axes in the t's space. It follows that for any integer p,

$$E\left(\sum_1^{n-1} z_iz_{i+1}\right)^p = E(I^p)\,E\left(\sum_1^n z_i^2\right)^p,$$

with a similar result for r_{c1}. To find $E(I^3)$ we have therefore only to find the expectations of the numerator and the denominator of (6). The expectation of the denominator is the third moment of χ^2 about zero with $(n-1)$ degrees of freedom and is therefore (Kendall, 1943, p. 292)

$$(n-1)(n+1)(n+3).$$

* It should be pointed out that the formula for k_4 given in Kendall (1943, p. 260) and in Dressel (1940, p. 45) is incorrect, in that the coefficient of the term s_4s_2 should be $-15n(n-1)^2(n+4)$. The correct value is given by Fisher (1929, p. 204).

To find the expectation of the numerator we must consider expressions such as $E(z_1^3 z_2^3)$, $E(z_1^3 z_2^3 z_3)$, etc., where the correlations between any two z's is $\rho = -(n-1)^{-1}$ and the standard deviation of any z is $\sigma = ((n-1)/n)^{\frac{1}{2}}$ if the standard deviation of the x's is unity. Isserlis (1918) has shown (and his results may be easily proved by using characteristic functions) that

$$E(z_1^3 z_2^3) = \sigma^6(9\rho + 6\rho^3), \qquad E(z_1^2 z_2 z_3 z_4 z_5) = \sigma^6(3\rho^2 + 12\rho^3),$$
$$E(z_1^3 z_2^2 z_3) = 3\sigma^6(\rho + 2\rho^2 + 2\rho^3), \quad E(z_1 z_2 z_3 z_4 z_5 z_6) = 15\sigma^6\rho^3.$$
$$E(z_1^2 z_2^2 z_3 z_4) = \sigma^6(\rho + 4\rho^2 + 10\rho^3),$$

Inserting these results in the expectation of the numerator of I^3 we get

$$\sigma^6\{(n-1)(9\rho + 6\rho^3) + 18(n-2)(\rho + 2\rho^2 + 2\rho^3) + 6(n-3)(\rho + 4\rho^2 + 10\rho^3)$$
$$+ 18(n-3)(n-4)(\rho^2 + 4\rho^3) + 3(n-2)(n-3)(\rho + 4\rho^2 + 10\rho^3) + 15(n-3)(n-4)(n-5)\rho^3\},$$

and using the values of σ and ρ we get

$$E(I^3) = -\frac{3(n^3 - 2n^2 + 2n - 5)}{n^3(n-1)(n+3)}.$$

Therefore

$$\mu_3'(r_1) = -\frac{3(n^3 - 2n^2 + 2n - 5)}{(n-1)^4(n+3)},$$

and

$$\mu_3(r_1) = \frac{4(n-4)(n-3)}{(n-1)^4(n+3)}, \tag{7}$$

a result which has been checked by the method previously described. For the cyclic definition

$$E(r_{c1}^3) = -\frac{3}{(n+3)(n-1)}$$

which agrees with the result of Dixon (1944), so that

$$\mu_3(r_{c1}) = \frac{2(2n-1)(n-6)}{(n-1)^3(n+1)(n+3)}. \tag{8}$$

(7) and (8) check with the results obtained by using the method previously described. In a similar way $\mu_4(r_1)$ and $\mu_4(r_{c1})$ could be obtained by somewhat more lengthy algebra. In the case of the cyclic definition, formulae for the moments of any order have been obtained by Dixon (1944) by using a generating function. However, his method, which is more complicated than the above, does not seem to apply easily when the serial correlation coefficient is defined by (1).

In practice we rarely know the true mean of the time series and so it is the distribution of the serial correlation coefficient corrected for the mean which we need to consider. However, it is worth pointing out that if the true mean is known, the first four moments of the serial correlation coefficient can be found very easily.

Consider first the analogue of Orcutt's definition where we know that the true mean is zero.

$$r_1 = \frac{n}{n-1} \frac{\sum_1^{n-1} x_i x_{i+1}}{\sum_1^n x_i^2}. \tag{9}$$

Then the denominator of (9) is distributed as χ^2 with n degrees of freedom and

$$E\left(\sum_1^n x_i^2\right)^p = n(n+2)\ldots(n+2p-2),$$

if we take, as we may, var (x_i) to be equal to unity. Then, arguing as before, we see that r_1 depends only on the angles which the vector $(x_1 \dots x_n)$ makes with the axes in the n-space of the x's and therefore, for any integer p,

$$E(r_1^p) = \left(\frac{n}{n-1}\right)^p \frac{E\left(\sum_1^{n-1} x_i x_{i+1}\right)^p}{n(n+2)\dots(n+2p-2)}.$$

Since the x's are independent we have

$$E\left(\sum_1^{n-1} x_i x_{i+1}\right) = 0, \quad E\left(\sum_1^{n-1} x_i x_{i+1}\right)^2 = E\left(\sum_1^{n-1} x_i^2 x_{i+1}^2\right) = n-1,$$

other terms in the expansion having zero expectation,

$$E\left(\sum_1^{n-1} x_i x_{i+1}\right)^3 = 0,$$

$$E\left(\sum_1^{n-1} x_i x_{i+1}\right)^4 = E\left\{\sum_1^{n-1} x_i^4 x_{i+1}^4 + 6 \sum_1^{n-2} x_i^2 x_{i+1}^4 x_{i+2}^2 + 3 \sum x_i^2 x_{i+1}^2 x_j^2 x_{j+1}^2\right\},$$

where the last sum is taken over distinct values of the suffixes $i, i+1, j, j+1$,

$$= 9(n-1) + 18(n-2) + 3(n-2)(n-3)$$

$$= 3(n^2 + 4n - 9),$$

and so

$$\left.\begin{aligned} &\mu_1'(r_1) = \mu_3(r_1) = 0, \\ &\mu_2(r_1) = \frac{n-1}{n(n+2)}\left(\frac{n}{n-1}\right)^2 = \frac{n}{(n-1)(n+2)}, \\ &\mu_4(r_1) = \frac{3n^3(n^2+4n-9)}{(n-1)^4(n+2)(n+4)(n+6)}. \end{aligned}\right\} \tag{10}$$

In a similar way, if we use the circular definition uncorrected for the mean,

$$r_{c1} = \frac{\sum_1^n x_i x_{i+1}}{\sum_1^n x_i^2}, \tag{11}$$

we find

$$\mu_1'(r_{c1}) = \mu_3(r_{c1}) = 0,$$

$$\mu_2(r_{c1}) = (n+2)^{-1} \quad (n \geqslant 5),$$

$$\mu_4(r_{c1}) = 3\{(n+2)(n+4)\}^{-1} \quad (n \geqslant 1),$$

which agree with Dixon's results. He gives a general formula for the moments of (11), but, as before, his method does not seem to apply to the non-circular case. The above method can be applied to find the moments of the distribution of such quantities as

$$\sum_1^{n-1} (x_{i+1} - x_i)^2 \Big/ \sum_1^n (x_i - \bar{x})^2,$$

which have been discussed by v. Neumann and others (see references in Dixon's paper). Unfortunately, as soon as there exists serial correlation in the population series, these

methods can no longer be applied. Madow (1945), however, has obtained the exact distribution of the cyclic serial correlation coefficient from a correlated universe as an extension of the results obtained by other writers for samples from uncorrelated universes (for references see Rubin (1945)).

REFERENCES

ANDERSON, R. L. (1942). *Ann. Math. Statist.* **13**, 1.
DIXON, W. J. (1944). *Ann. Math. Statist.* **15**, 119.
DRESSEL, P. L. (1940). *Ann. Math. Statist.* **11**, 33.
FISHER, R. A. (1929). *Proc. Lond. Math. Soc.* (2), **30**, 199.
FISHER, R. A. (1930). *Proc. Roy. Soc.* A, **130**, 16.
ISSERLIS, L. (1918). *Biometrika*, **12**, 134.
KENDALL, M. G. (1943). *The Advanced Theory of Statistics*, **1** London: Charles Griffin and Co.
KOOPMANS, T. (1942). *Ann. Math. Statist.* **13**, 14.
MADOW, W. G. (1945). *Ann. Math. Statist.* **16**, 308.
ORCUTT, G. H. (1948). *J. R. Statist. Soc. B*, **10**, 1.
RUBIN, H. (1945). *Ann. Math. Statist.* **16**, 211.

SOME RESULTS IN THE TESTING OF SERIAL CORRELATION COEFFICIENTS

By M. H. QUENOUILLE

Anderson (1942) and Koopmans (1942) have recently investigated the distribution of the serial correlation coefficient defined by

$$r_l = \frac{\sum_{i=1}^{n} x_i x_{i+l}}{\sum_{i=1}^{n} x_i^2},$$

where the x_i are normally and independently distributed about zero, and x_{n+i} is taken as x_i. This definition approximates to the definition

$$r_l = \frac{\sum_{i=1}^{n-l} x_i x_{i+l}}{\sqrt{\left(\sum_{i=1}^{n-l} x_i^2 \sum_{i=1}^{n-l} x_{i+l}^2\right)}}$$

which is more generally used, and using this form, Anderson was able to obtain an exact distribution. Anderson's distribution was, however, difficult to use in practice, and Koopmans obtained an integral approximation, which was adequate for $n \geqslant 10$. Rubin (1945) has solved this integral to give the distribution

$$h(r)\,dr = \frac{\Gamma(\frac{1}{2}n+1)}{\Gamma(\frac{1}{2}n+\frac{1}{2})\,\Gamma(\frac{1}{2})}(1-r^2)^{\frac{1}{2}(n-1)}\,dr, \tag{1}$$

which is, in fact, the distribution of the ordinary correlation coefficient based on $n+3$ observations. Madow (1945) has demonstrated how this distribution may be extended, and if the x_i are connected by a linear Markoff scheme $x_{i+1} = \rho x_i + \epsilon_{i+1}$ where the ϵ_i are normally and independently distributed about zero, it is not difficult to show that a good approximation to the distribution of the first serial correlation coefficient is given by

$$h(r)\,dr = \frac{\Gamma(\frac{1}{2}n+1)}{\Gamma(\frac{1}{2}n+\frac{1}{2})\,\Gamma(\frac{1}{2})}\frac{(1-r^2)^{\frac{1}{2}(n-1)}}{(1-2\rho r+\rho^2)^{\frac{1}{2}n}}. \tag{2}$$

We can, in a manner analogous to the method used with the ordinary correlation coefficient, make the transformation

$$r = \tanh z, \quad \rho = \tanh\zeta \quad \text{and} \quad x = z-\zeta \text{ in the form (2).}$$

Then $$h(x)\,dx = \frac{\Gamma(\frac{1}{2}n+1)}{\Gamma(\frac{1}{2}n+\frac{1}{2})\,\Gamma(\frac{1}{2})}\frac{dx}{\cosh\zeta\,\cosh^{n+1}x(1+\tanh\zeta\,\tanh x)\,(1-\tanh^2\zeta\,\tanh^2 x)^{\frac{1}{2}n}}.$$

Now, for n large, $x = O(1/\sqrt{n})$ and

$$\log_e \frac{\Gamma(\frac{1}{2}n+1)}{\Gamma(\frac{1}{2}n+\frac{1}{2})\,\Gamma(\frac{1}{2})} = \tfrac{1}{2}\log_e\frac{n}{2\pi} + \frac{1}{4n} + O\left(\frac{1}{n^3}\right),$$

$$\log_e\cosh x = \frac{x^2}{2!} - \frac{2x^4}{4!} + O\left(\frac{1}{n^3}\right),$$

$$\log_e(1-\rho^2\tanh^2 x) = -\rho^2\left(x^2 - \frac{2x^4}{3}\right) - \frac{\rho^4 x^4}{2} + O\left(\frac{1}{n^3}\right),$$

$$\log_e(1+\rho\tanh x) = \rho\left(x - \frac{x^3}{3}\right) - \frac{\rho^2 x^2}{2} + \frac{\rho^3 x^3}{3} + O\left(\frac{1}{n^2}\right),$$

whence
$$h(x) = \frac{1}{\sqrt{(2\pi)}\sigma}\exp\left\{-\frac{x^2}{2\sigma^2}\right\}X,$$

where
$$\sigma^2 = \frac{\cosh^2\zeta}{n} = \frac{1}{n(1-\rho^2)},$$

and
$$X = 1-\rho x+\left[\frac{1}{4n}-\frac{x^2(1-\rho^2)}{2}+\frac{nx^4(1-\rho^2)(1-3\rho^2)}{12}+\frac{\rho^2x^2}{2}\right]$$
$$-\rho x\left[\frac{1}{4n}-\frac{5x^2(1-\rho^2)}{6}+\frac{nx^4(1-\rho^2)(1-3\rho^2)}{12}+\frac{\rho^2x^2}{6}\right]+O\left(\frac{1}{n^2}\right).$$

We can now obtain the moments of the distribution in the form

$$\mu_1 = -\frac{\rho}{n(1-\rho^2)}+\frac{\rho(1+\rho^2)}{n^2(1-\rho^2)^2}+O\left(\frac{1}{n^3}\right), \quad \mu_2 = \frac{1}{n(1-\rho^2)}-\frac{2\rho^2}{n^2(1-\rho^2)^2}+O\left(\frac{1}{n^3}\right),$$
$$\mu_3 = \frac{6\rho^3}{n^3(1-\rho^2)^3}+O\left(\frac{1}{n^4}\right), \qquad \mu_4 = \frac{3}{n^2(1-\rho^2)^2}+\frac{2(1-9\rho^2)}{n^3(1-\rho^2)^3}+O\left(\frac{1}{n^4}\right),$$

and the coefficients of skewness and kurtosis are given by

$$\gamma_1 = \frac{6\rho^3}{[n(1-\rho^2)]^{\frac{3}{2}}}+O\left(\frac{1}{n^{\frac{3}{2}}}\right), \quad \gamma_2 = \frac{2(1-3\rho^2)}{n(1-\rho^2)}+O\left(\frac{1}{n^2}\right).$$

Hence, under this transformation, z will be distributed approximately normally about the mean
$$\zeta-\frac{\rho}{n(1-\rho^2)}+\frac{\rho(1+\rho^2)}{n^2(1-\rho^2)^2},$$

with variance
$$\frac{1}{n(1-\rho^2)}-\frac{2\rho^2}{n^2(1-\rho^2)^2}.$$

It should be noted that, compared with the transformation of the ordinary correlation coefficient, this transformation has the disadvantage that the variance of z is, to the first order, dependent upon the value of ρ. Furthermore, for $|\rho|$ large, the mean value of z deviates more widely from ζ.

The following examples are intended to investigate and demonstrate the testing of serial correlation coefficients when the error distribution of the ϵ is not normal.

In these examples, 1, 4 and 5 test the approximation by sampling artificial series with rectangular error distributions, 2 and 3 use an exceptional error distribution to show that the approximation is valid for n as low as twenty, while examples 6 and 7 demonstrate that the values of r obtained from certain observed series are consistent with homogeneity of the data.

Example 1. Serial correlation coefficients were calculated (according to the circular definition) from twenty sets of twenty random numbers, rectangularly distributed from -49 to 49.

The values of r thus obtained were:

−0·348,	−0·257,	−0·211,	−0·202,	−0·200,	−0·174,	−0·137,
−0·067,	−0·058,	−0·055,	−0·030,	0·011,	0·100,	0·112,
0·177,	0·234,	0·236,	0·348,	0·376,	0·445.	

These values were transformed by the transformation $r = \tanh z$ so that z should be distributed with zero mean, and variance 0·0500. The estimated mean and variance about zero were found to be 0·017 and 0·0530 respectively. The normality was, in addition, tested by the calculation of skewness and kurtosis coefficients

$$g_1 = 0{\cdot}595 \pm 0{\cdot}512, \quad g_2 = -0{\cdot}751 \pm 0{\cdot}992.$$

For small samples such as this, it must be borne in mind that the values of g_1 and g_2 do not demonstrate normality but only the absence of extreme non-normality.

Example 2. Suppose we consider the distribution $P(x_i = 1) = P(x_i = -1) = \frac{1}{2}$, so that $E(x_i x_{i+1}) = 0$. Then each product of successive observations will be distributed in the same manner as x_i, and r, based on n such products, which are independent, is distributed binomially (if a circular definition is employed, then the same distribution of r, omitting alternate ordinates, is obtained). Thus we can compare the terms of a binomial of degree n with the ordinates of the distribution of the ordinary correlation coefficient based on $n+3$ observations. The values of these are given below:

$n = 5$

r	Binomial terms	Correlation coefficient ordinates	Binomial terms $\div$ 0·4
1·0	0·031	0·00	0·08
0·6	0·156	0·38	0·39
0·2	0·312	0·86	0·78

$n = 10$

r	Binomial terms	Correlation coefficient ordinates	Binomial terms $\div$ 0·2
1·0	0·001	0·00	0·00
0·8	0·010	0·01	0·05
0·6	0·044	0·17	0·22
0·4	0·117	0·59	0·59
0·2	0·205	1·08	1·03
0·0	0·246	1·29	1·23

$n = 20$

r	Binomial terms	Correlation coefficient ordinates	Binomial terms $\div$ 0·1
0·8	0·000	0·00	0·00
0·7	0·001	0·00	0·01
0·6	0·005	0·03	0·05
0·5	0·015	0·12	0·15
0·4	0·037	0·35	0·37
0·3	0·074	0·74	0·74
0·2	0·120	1·23	1·20
0·1	0·160	1·64	1·60
0·0	0·176	1·81	1·76

The comparison of the discrete and continuous distributions is difficult, but it is clear that, in view of the nature of the error distribution, the approximation is good for n as low as 20.

Example 3. The method of Example 2 can be used for the case when $\rho \neq 0$. If we consider the distribution

$$P(x_{i+1} = 1 | x_i = 1) = P(x_{i+1} = -1 | x_i = -1) = p,$$
$$P(x_{i+1} = 1 | x_i = -1) = P(x_{i+1} = -1 | x_i = 1) = q.$$

Then $E(x_i x_{i+1}) = (p-q)^l$, and this will behave similar to a linear Markoff scheme with $\rho = p-q$, except that only the values ± 1 are admissible. In this case, r is distributed binomially with parameter p so that we can compare the discrete binomial with the continuous theoretical distribution, as below. In this case, the fit is less satisfactory, since the binomial distribution is less skew, but, considering the approximate nature of the assumption, the fit is remarkable for n as low as 20.

$n = 20$

r	$p=\frac{3}{4}$ 10 (binomial terms)	$\rho=\frac{1}{2}$ Correlation coefficient ordinates	$p=\frac{3}{4}$ 10 (binomial terms)	$\rho=\frac{1}{2}$ Correlation coefficient ordinates
1·0	0·00	0·00	0·03	0·00
0·9	0·03	0·00	0·21	0·05
0·8	0·14	0·03	0·67	0·30
0·7	0·42	0·24	1·34	1·20
0·6	0·91	0·79	1·89	1·93
0·5	1·46	1·45	2·03	2·09
0·4	1·82	1·87	1·69	1·75
0·3	1·82	1·87	1·12	1·23
0·2	1·48	1·54	0·61	0·75
0·1	0·99	1·06	0·27	0·40
0·0	0·54	0·63	0·10	0·19
−0·1	0·25	0·32	0·03	0·08
−0·2	0·09	0·14	0·01	0·03
−0·3	0·03	0·05	0·00	0·01
−0·4	0·01	0·01	—	—
−0·5	0·00	0·00	—	—

Example 4. Serial correlation coefficients were calculated, using the ordinary definition, from thirty sets of twenty-one serial correlated numbers. These numbers had been derived using the scheme $x_{i+1} = 0{\cdot}5x_i + \epsilon_{i+1}$, where the ϵ_i were rectangularly distributed from -49 to 49. The thirty values of r thus obtained were

−0·113, 0·275, 0·282, 0·327, 0·339, 0·352, 0·386, 0·408, 0·411, 0·424,
0·428, 0·435, 0·445, 0·446, 0·476, 0·518, 0·547, 0·550, 0·562, 0·577,
0·590, 0·613, 0·614, 0·630, 0·645, 0·649, 0·672, 0·677, 0·718, 0·744.

These values were then transformed, and, in this case, we expect z to be distributed about mean

$$0{\cdot}5494 - \frac{0{\cdot}5}{20(0{\cdot}75)} + \frac{(0{\cdot}5)(1{\cdot}25)}{400(0{\cdot}75)^2} = 0{\cdot}5209,$$

with variance

$$\frac{1}{20(0{\cdot}75)} - \frac{0{\cdot}5}{400(0{\cdot}75)^2} = 0{\cdot}0644.$$

The estimated values of the mean and of the variance about the theoretical mean were 0·557 and 0·0517, while $g_1 = 0{\cdot}024 \pm 0{\cdot}427$ and $g_2 = 0{\cdot}785 \pm 0{\cdot}833$ indicated no significant deviation from normality.

Example 5. M. G. Kendall (1946) has calculated correlograms for eight sub-series of a series of 480 terms of the autoregressive scheme $x_{i+2} = 1{\cdot}1x_{i+1} - 0{\cdot}5x_i + \epsilon'_{i+2}$, where the ϵ'_i are rectangularly distributed from -49 to 49.

We might test his eight values of r_1 against the theoretical value $\rho_1 = 0{\cdot}733$. In this case, the ϵ_i in the scheme $x_{i+1} = \rho_1 x_i + \epsilon_{i+1}$ are correlated. The expected values of the mean and variance of the z distribution are easily found to be 0·9099 and 0·0353, and these may be compared with the mean, 1·0181, and the variance about the theoretical mean, 0·0454, calculated from the transformed values, 0·7498, 0·9181, 0·9223, 0·9439, 0·9962, 1·0352, 1·1786, 1·4007. The normality can again be tested giving

$$g_1 = 0{\cdot}567 \pm 0{\cdot}752, \quad g_2 = 0{\cdot}876 \pm 1{\cdot}481.$$

Example 6. Sir Gilbert Walker (1946) has calculated serial correlation coefficients for pressure data observed at daily intervals. His values for the period October to March in the years 1930–6 were 0·76, 0·88, 0·80, 0·88 0·79 and 0·82. These can be transformed and tested giving $s^2 = 0{\cdot}0215$ compared with the theoretical $\sigma^2 = 0{\cdot}0174$ approximately.

A similar set of unpublished data calculated for six successive months gave values of r_1 equal to 0·62, 0·76, 0·86, 0·78, 0·60, 0·86 with $s^2 = 0{\cdot}0688$ and $\sigma^2 = 0{\cdot}0803$ approximately.

Example 7. The Beveridge series of trend-free wheat-price index numbers was split into sixteen subseries of twenty-three terms each, and serial correlation coefficients were calculated for these subseries. The values obtained were 0·379, 0·460, 0·329, 0·421, 0·772, 0·464, 0·678, 0·767, 0·728, 0·311, 0·497, 0·656, 0·270, 0·508, 0·691 and 0·589 respectively. These were transformed, and the mean and variance of the transformed values were found to be 0·625 and 0·0625. If we assume $\rho = \tanh 0{\cdot}625 = 0{\cdot}5545$, then

$$\sigma^2 = \frac{1}{22(0{\cdot}6925)} - \frac{0{\cdot}6150}{484(0{\cdot}6925)^2} = 0{\cdot}0631,$$

which agrees remarkably well with the observed variance.

SUMMARY

We may summarize the results of these examples in tabular form, as below:

Example	ρ	σ^2	s^2	Degrees of freedom	χ^2	P
1	0·0	0·0500	0·0530	20	21·20	0·39
4	0·5	0·0644	0·0517	30	24·08	0·77
5	0·733	0·0353	0·0454	8	10·29	0·25
6 (*a*)	0·827	0·0174	0·0215	5	6·18	0·29
(*b*)	0·765	0·0803	0·0688	5	4·28	0·51
7	0·554	0·0631	0·0625	15	14·86	0·46
Total				83	80·89	0·55

Thus, while the results given here do not constitute definite proof, there is every indication that the approximate normal theory provides a satisfactory test for serial correlation coefficients, if the number of observations is sufficiently high, and such a test is undoubtedly

useful in numerous ways, varying from the determination of density of observations in sampling enquiries to the comparison of 'after-effects' in biological or economic research.

It is worth noting that the method of Examples 2 and 3 can be used to investigate the approximate test commonly used to test the correlation between two serially correlated series. If the correlation between successive terms of the two series are ρ_1 and ρ_2, then the correlation between corresponding terms of the series is usually tested with

$$n' = n(1-\rho_1\rho_2)/(1+\rho_1\rho_2)$$

degrees of freedom, i.e. as if the correlation were based on n' pairs of observations. This result was first proved by Bartlett (1935).

Suppose we consider two schemes of the kind described in Example 3, with parameters $\rho_1 = p_1 - q_1$ and $\rho_2 = p_2 - q_2$. Then if

$$p = p_1p_2 + q_1q_2 = \tfrac{1}{2}(1+\rho_1\rho_2), \quad q = p_1q_2 + p_2q_1 = \tfrac{1}{2}(1-\rho_1\rho_2) \quad \text{and} \quad \rho = p - q = \rho_1\rho_2,$$

it is not difficult to show that the distribution of the correlation r between the two schemes is given by

$$P(r=1) = P(r=-1) = p^{n-1},$$

$$P\left(r = 1-\frac{2i}{n}\right) = p^{n-2}q + (\tfrac{1}{2}n-1)p^{n-3}q^2 + [(\tfrac{1}{2}n-1)^2 - r^2(\tfrac{1}{2}n)^2]p^{n-4}q^3$$

$$+ \frac{1}{1!\,2!}(\tfrac{1}{2}n-2)[(\tfrac{1}{2}n-1)^2 - r^2(\tfrac{1}{2}n)^2]p^{n-5}q^4$$

$$+ \frac{1}{(2!)^2}[(\tfrac{1}{2}n-2)^2 - r^2(\tfrac{1}{2}n)^2][(\tfrac{1}{2}n-1)^2 - r^2(\tfrac{1}{2}n)^2]p^{n-6}q^5 + \ldots.$$

The ordinates of this distribution are given below, for $n = 20$ and

$\rho = 1{\cdot}0, \quad 0{\cdot}8, \quad 0{\cdot}5, \quad 0{\cdot}0, \quad -0{\cdot}5, \quad -0{\cdot}8, \quad 1{\cdot}0,$

$p = 1{\cdot}0, \quad 0{\cdot}9, \quad 0{\cdot}75, \quad 0{\cdot}5, \quad 0{\cdot}25, \quad 0{\cdot}1, \quad 0{\cdot}0,$

$n' = 0, \quad 2{\cdot}2, \quad 6{\cdot}7, \quad 20{\cdot}0, \quad 60{\cdot}0, \quad 180{\cdot}0, \quad \infty.$

r \ ρ	1·0	0·8	0·5	0	−0·5	−0·8	−1·0
1·0	0·5	0·0675	0·0021	—	—	—	—
0·9	0·0	0·0300	0·0056	—	—	—	—
0·8	0·0	0·0346	0·0118	0·0002	—	—	—
0·7	0·0	0·0389	0·0210	0·0011	—	—	—
0·6	0·0	0·0430	0·0332	0·0046	—	—	—
0·5	0·0	0·0466	0·0475	0·0148	0·0004	—	—
0·4	0·0	0·0498	0·0629	0·0370	0·0036	—	—
0·3	0·0	0·0523	0·0776	0·0739	0·0238	0·0013	—
0·2	0·0	0·0542	0·0898	0·1201	0·0954	0·0263	—
0·1	0·0	0·0553	0·0980	0·1602	0·2256	0·2219	—
0·0	0·0	0·0557	0·1008	0·1762	0·3025	0·5010	1·0000

A comparison of these values with tables of the distribution of the correlation coefficient constructed by David (1938) shows that they are distributed to a high degree of approxima-

tion with $n'+2$ degrees of freedom. The extra two degrees of freedom are more likely due to the nature of the error distribution than to the approximate nature of n', since they are independent of ρ. In any case, the conclusion reached is that the approximation is valid provided that the *effective* number of degrees of freedom is large.

REFERENCES

ANDERSON, R. L. (1942). *Ann. Math. Statist.* **13**, 1.

BARTLETT, M. S. (1935). *J. R. Statist. Soc.* **98**, 536.

DAVID, F. N. (1938). *Tables of the Ordinates and Probability Integral of the Distribution of the Correlation Coefficient in Small Samples.* Cambridge University Press.

KENDALL, M. G. (1946). *Contributions to the Study of Oscillatory Time Series.* National Institute of Economic and Social Research.

KOOPMANS, T. (1942). *Ann. Math. Statist.* **13**, 14.

MADOW, W. G. (1945). *Ann. Math. Statist.* **16**, 308.

RUBIN, H. (1945). *Ann. Math. Statist.* **16**, 211.

WALKER, Sir G. (1946). *Quart. J.R. Met. Soc.* **72**, 265.

FRACTIONAL REPLICATION ARRANGEMENTS FOR FACTORIAL EXPERIMENTS WITH FACTORS AT TWO LEVELS

By K. A. BROWNLEE, B. K. KELLY and P. K. LORAINE
The Research Department of the Distillers Company Ltd., Great Burgh, Epsom, Surrey

The theory of fractional replication of factorial experiments has been given recently (Finney, 1945; Finney, 1946; Kempthorne, 1947). The present paper is a description of the solutions of practical value in the case of the 2^{n-m} experiment (that is, the $(n-m)$th replicate of n factors all at two levels, involving the use of 2^m plots). Such confounding and double confounding arrangements as have been found will be given.

Two useful solutions are also included where one and two factors respectively are at four levels, these being derived from solutions with all factors at two levels.

It will be assumed that the reader is familiar with the basic concepts of fractional replication as set out by Finney (1945). These briefly are the group properties of treatment symbols and effect symbols, the use of generators to define concisely groups and subgroups, the concept of orthogonality between two subgroups, and the use of high-order interactions as aliases of main effects and first-order interactions.

In confounding with fractional replication the confounding subgroup with its aliases cannot contain any element in common with the alias subgroup except the identity and for practical purposes should not contain main effects nor, as far as possible, first-order interactions.

Fisher (1942*a*) introduced double confounding in which the two confounding subgroups must not contain any element other than the identity in common. In fractional replication there is the additional restriction that there should be no overlapping of the aliases.

To obtain the actual arrangement of plots in any particular simply confounded experiment, having selected the alias subgroup and the confounding subgroup, we form the so-called 'principal block' from those treatments whose symbols are orthogonal to both these subgroups. Subsequent blocks can be obtained by multiplying the principal block by symbols orthogonal to the alias subgroup which have not previously occurred. With double confounding these multipliers are of course the treatments of the principal block corresponding to the second confounding subgroup.

Enumeration of subgroups

Kempthorne (1947) reports that no simple method of enumerating subgroups suitable for high-order fractional replication has been found. It seems worth while, therefore, to record a method which has been found satisfactory for all designs of practical interest. It inevitably becomes more laborious for higher order subgroups, but such an enumeration is likely to be of sufficient use to justify the time spent in arriving at it.

Subgroups of order 2^m of the effects group involving n factors can be subdivided into types, two subgroups being considered of the same type when a permutation of the n letters representing the factors converts one into the other. Each type may be represented by a symbol

$$(n_1, n_2, n_3, \ldots),$$

where the $(2^m - 1)$ numbers n_1, n_2, etc., are the numbers of letters appearing in the separate elements of subgroups of that type. The symbolism may be further condensed by attaching to each number a suffix indicating how many times it is repeated. Thus the alias subgroup given below for the $\frac{1}{16}$ replicate of 8 factors would be represented by the symbol

$$(4_{14}, 8).$$

This symbolism is not perfect, as examples do occur where the same symbol represents more than one type. In such cases, the separate types will be distinguished by suffices outside the brackets.

The following conditions are *necessary* for a symbol to represent a type of subgroup:

(1) The sum of all the numbers in the brackets must equal 2^{m-1} times the number of letters actually used in the type.

(2) The numbers in the brackets must be all even or 2^{m-1} of them must be odd.

(3) When 2^{m-1} of the numbers are odd, the even numbers must, with the identity, form a subgroup of order 2^{m-1}.

(4) If the number n appears, the remaining numbers must be divisible into pairs such that the total of each pair is n.

(5) If the number 1 appears, the remaining numbers must be divisible into pairs such that the numbers in each pair differ by 1.

No other necessary conditions appear susceptible to expression in simple general terms.

The actual process of enumeration is best demonstrated by an example. Suppose it is required to find all the subgroups of order 8 involving 7 letters. The first step is to find all the combinations of 7, i.e. $2^m - 1$, even numbers which total to 28, i.e. $n \times 2^{m-1}$. Each of these is taken in turn and 2, i.e. 2^{m-2}, of the numbers are increased by 1 and the same number reduced by 1 in every possible way, bearing in mind the total number of elements containing a particular number of letters in the complete group. Next any duplication introduced by this process (some symbols will be derived from more than one combination of even numbers) is eliminated. The result will be all the symbols satisfying conditions (1) and (2) above. Further symbols are then rejected through failure to satisfy conditions (3), (4) and (5) above.

The final step consists in the examination of the remaining symbols for the existence of corresponding subgroups. This is best carried out by working from the even subgroup of order 2^{m-1} contained within it. Any one of the odd numbers may be taken for the last generator, and the number of factors this element must have in common with each of the even elements can be determined by the fact that the products must give the other odd elements. It can then be readily determined whether such a generator exists. The all-even symbols may be examined in a similar way except that there is a wider choice for the subgroup of order 2^{m-1}. At this stage, symbols representing more than one type of subgroup must be watched for. Considerable assistance can be obtained here by carrying out a check of the total number of subgroups.

The total number of subgroups of order 2^m in the group of order 2^n is given by (Carmichael, 1937)

$$\frac{(2^n - 1)(2^n - 2)(2^n - 2^2) \ldots (2^n - 2^{m-1})}{(2^m - 1)(2^m - 2)(2^m - 2^2) \ldots (2^m - 2^{m-1})}.$$

Some of these subgroups, however, do not use all the n factors. The number using all factors can be determined by a calculation exemplified in Table 1. The second column is derived by

applying the formula given by Carmichael. The third column is derived from the second column by subtracting $\binom{n}{3}$ from each entry. The fourth column is derived from the third by subtracting 11 times $\binom{n}{4}$ from each item and so on until the single entry in the sixth column is obtained. This gives the number of subgroups of order 8 which employ all the 7 factors, 7000 such subgroups employing 6 factors or less.

Table 1

(1) Factors n	(2) Total subgroups of order 8	(3)	(4)	(5)	(6)
3	1	—	—	—	—
4	15	11	—	—	—
5	155	145	90	—	—
6	1,395	1,375	1,210	670	—
7	11,811	11,776	11,391	9,501	4,811

When the existence of a subgroup corresponding to any symbol has been established, the number of subgroups of this type can be found by solving the appropriate combinatorial problem after one member of the type, or merely a set of generators, has been written out in full. Failure of the total to agree with that found as above after all symbols have been dealt with indicates, in the absence of errors, that at least one symbol represents more than one type of subgroup. In searching for this the all-even symbols should be examined first as these are usually, if not always, the source of the trouble.

If only subgroups containing no elements with less than five letters are considered of interest, the work can be considerably lightened by carrying out the above process but only starting from the all-even combinations which contain no numbers less than 4 and, in enumerating from these, only considering symbols in which all the 4's are raised to 5.

In this case, however, there does not seem to be any numerical check, so that there is some risk of missing one or more types through their being represented by the same symbols as other types. The risk is not very serious, however, since types represented by the same symbol have most properties which are important from the point of view of fractional replication in common.

Designs based on sixteen plots

There are designs based on 16 plots for 6, 7, or 8 factors ($\frac{1}{4}$, $\frac{1}{8}$, and $\frac{1}{16}$ replicates respectively), which give main effects with second-order interactions as aliases. First-order interactions must be used as error, so if these exist and are large the residual will be inflated accordingly.

The most general design is that for 8 factors. It is based on the alias subgroup of the type $(4_{14}, 8)$ generated by $ABCD$, $CDEF$, $ACEG$, $EFGH$.

This can be confounded in 4 blocks of 4, confounding AB, BC and AC, with

(1), *abcd*, *efgh*, *abcdefgh*,

as the first block and subsequent blocks being obtained by multiplying this block successively by *abef*, *aceg*, *adeh*.

The designs for 6 and 7 factors can be obtained from the above. The alias subgroups for these two cases are obtained by deleting those generators containing letters beyond the sixth and seventh respectively, and the plot treatments are obtained by deleting the letters beyond the sixth and seventh respectively where they occur.

DESIGNS BASED ON THIRTY-TWO PLOTS

The two extreme designs based on 32 plots are the half-replicate of 6 factors, which gives its main effects as aliases of fourth-order interactions and its first-order interactions as aliases of third-order interactions, and the $\frac{1}{32}$ replicate of 10 factors, which gives its main effects as aliases of second-order interactions and no first-order interactions. Between these two extremes come designs in which the order of aliases falls as the degree of completeness of replication falls.

The possibilities are set out in Table 2. The confounding arrangements for 6 and 7 factors can be used for double confounding, since in each case the two confounding subgroups and their aliases contain no element in common other than the identity. However, in the case of 7 factors, where it is desired merely simply to confound in 4 blocks of 8, it is better to use the confounding subgroup I, ACF, ACG, FG as FG is already lost, so nothing further is lost by this restriction.

It will be noted that the alias subgroup for the $\frac{1}{16}$ replicate has the structure $(4_7, 5_7, 9)$. At first sight one would have expected the subgroup generated by $ABCD$, $ABEF$, $ABGH$, $ACEGJ$ with the structure $(4_6, 5_8, 8)$ to be more satisfactory, but in point of fact, with this alias subgroup, only one factor retains its first-order interactions.

DESIGNS BASED ON SIXTY-FOUR PLOTS

The designs for 64 plots proceed from the half-replicate of 7 factors with a high degree of security in its aliases, first-order interactions having fourth-order interactions as aliases, to the $\frac{1}{32}$ replicate of 11 factors with all but three main effects having only second-order interactions as aliases.

The possibilities are shown in Table 3. With 8 blocks of 8, in all cases double confounding is possible with zero or slight loss of first-order interactions. With 16 blocks of 4 and 4 blocks of 16, satisfactory doubly confounded arrangements have been found only for 7, 8 and 11 factors, but simply confounded arrangements are given for 9 and 10 factors.

DESIGNS WITH MORE THAN SIXTY-FOUR PLOTS

For experiments with more than 64 plots the investment in most fields of experimentation is so great that it becomes essential to have a higher degree of certainty in the results, and hence the degree of ambiguity which can be tolerated with respect to the aliases is lower, i.e. the order of the interactions forming the aliases must be relatively higher.

The possibilities are set out in Table 4. The alias subgroup for 12 factors is due to Fisher (1942*b*). In all these designs the aliases of main effects are fourth-order interactions or better, and the aliases of first-order interactions are third-order interactions or better. The confounding arrangements given do not lose any first-order interactions.

There are certain less secure designs which may prove useful in certain circumstances and these are briefly summarized in Table 5.

Table 2. *Designs based on 32 plots*

No. of factors	6	7	8	9	10
Order of replication	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$
Generators of alias subgroup	*ABCDEF*	*ABCDE* *CDFG*	*ABCD* *ABEF* *ACFGH*	*ABCD* *ABEF* *ACEG* *ABGHJ*	*ABCD* *CDEF* *EFGH* *GHJK* *ACEGJ*
Main effects having as aliases interactions of order: 4	All	—	—	—	—
3		*A, B, E*	*G, H*	*H, J*	—
2		*C, D, F, G*	*A, B, C, D, E, F*	*A* to *G*	All
First orders retained as third orders	All	0	0	0	0
Factors retaining all first-order interactions as second orders	—	*A, B, E*	*G, H*	*H, J*	None
No. of such interactions	—	15	13	15	0
No. of other first orders with second orders as aliases	—	0	0	0	0
No. of first orders lost	0	6	15	21	All
Confounding in 4 blocks of 8 and 8 blocks of 4:		Double		Single	
Generators of confounding subgroups	*ABC* *AEF*	*ACE* *BDE* *BCF*	*CG* *BCD*	*ACF* *FG* *BG*	*CD* *DF*	*CD* *DF* *AD*	*AG* *EG*	*AG* *EG* *AF*	*CD* *DF*	*CD* *DF* *BC*
Further first-order interactions lost other than those contained in the confounding subgroup	*AD*	*AB* *CD* *EF*	*AE*	*CE* *DE*	—	—	—	—	—	—
Total no. of new first-order interactions lost by confounding	1	3	1	4	0	0	0	0	0	0
Generators of principal blocks	*ef* *bc* *acde*	*abcd* *abef*	*ae* *bcf* *bdf*	*acde* *abfg*	*gh* *abg* *cdef*	*gh* *abcdefg*	*hj* *aceg* *bdfh*	*hj* *abcdefgh*	*cdef* *abgh* *ghjk*	*ghjk* *abcdefgh*

No. of factors	7	8	9	10	11

No. of factors	7	8	9	10	11
Order of replication	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$
Generators of alias subgroup	*ABCDEFG*	*ABCDE* *ABFGH*	*ABCD* *ABEFG* *ACEHJ*	*ADHJ* *BEGK* *ABCGH* *ABFJK*	*ABCD* *CDEF* *EFGH* *GHJKL* *ACEGJ*
Main effects having as aliases interactions of order: 5	All	—	—	—	—
4	—	—	—	—	—
3	—	All	*E, F, G, H, J*	*C, F*	*J, K, L*
2	—	—	*A, B, C, D*	*A, B, D, E, G, H, J, K*	*A* to *H*
No. of first orders retained as fourth orders	All	0	0	0	0
No. of first orders retained as third orders	—	9	0	0	0
Factors retaining all first-order interactions as second orders	—	*A, B*	*E* to *J*	*C, F*	*J, K, L*
No. of such interactions	—	13	30	17	27
No. of other first orders with second orders as aliases	—	6	0	16	0
No. of first orders completely lost	0	0	6	12	28
Double confounding in 8 blocks of 8:					
Generators of confounding subgroups	*ABC* *ABD* *CDE* *DEF* *BDF* *BFG*	*AB* *AH* *ADF* *ADG* *ACG* *BEF*	*BC* *FJ* *CD* *AH* *BE* *HJ*	*AB* *AF* *BC* *AG* *AJ* *BE*	*AB* *AF* *BC* *JK* *AJ* *AG*
Further first-order interactions lost other than those contained in the confounding subgroups	— — — —	*EH* *EG* *BC*	*AE* *AG*	— *CD* *FK*	*DJ* —
Total no. of new first-order interactions lost by confounding	0 0	2 3	4 7	5 5	4 1
Generators of principal blocks	*acdf* *bcdf* *bcef* *adfg* *bcdg* *abef*	*cegh* *defg* *abcd* *bcf* *cdfg* *adh*	*abcdefh* *bce* *fghj* *bdg* *hj* *acfhj*	*degh* *cdh* *efk* *behj* *abcdhjk* *afgjk*	*efgh* *abefghjk* *kl* *bcehl* *abcdghjl* *cdjk*
	Double		Single		Double
Confounding in 16 blocks of 4 and 4 blocks of 16: Generators of confounding subgroups	*AB* *ABCD* *BCF* *ABEG* *CDG* *DE*	*AB* *CDF* *BCF* *ADEH* *CDG* *DEH*	*AE* *BC* *FJ* *CD* *AH* *BC*	*AB* *AD* *BC* *DH* *BE* *BJ*	*AF* *AB* *JK* *BC* *AG* *AK*
Further first-order interactions lost other than those contained in the confounding subgroups	*AG* *BG* *EF* — *DF*	*AE* *BE* *EG* — *BG* *DF* *AG* *CH*	*BE* *DE* *DG* — *DH* *EG* *GH*	*DF* —	*CL* —
Total no. of new first-order interactions lost by confounding	6 0	8 0	10 0	9 0	8 0
Generators of principal blocks	*cdef* *ab* *abcg* *adef* *eg* *cd*	*abdefg* *defg* *cdfh* *abcd* *abgh* *bcf*	*abcdegh* *abcd* *bcfj* *hj* *fg* *efh*	*abcegjk* *bcef* *dfghk* *adhj* *bck* *bfg*	*bcehl* *abcd* *abefghjk* *ghjk* *efgh* *kl*

Table 4. *Designs with more than 64 plots*

No. of factors	9	12	16
Order of replication	$\frac{1}{4}$	$\frac{1}{16}$	$\frac{1}{128}$
No. of plots	128	256	512
Generators of alias subgroup	*ABCDEF* *DEFGHJ* — — — — —	*ABEFLM* *ABGHJK* *ACEGKM* *ADFGJM* — — —	*ABCDEFGH* *EFGHJKLM* *ABEFJKNO* *ACEGJLPQ* *ADEHJMNP* *ADEGNQ* *ACJMNO*
Confounding in 8 blocks: Generators of confounding subgroup	*ADG* *BEH* *CFJ*	*ABE* *CDE* *ACG*	*ACE* *BDE* *ADF*
Generators of principal block	*acdf* *abde* *degh* *efhj* — —	*abcd* *bcegjk* *acehlm* *jklm* *fhjl* —	*abcd* *jklm* *nopq* *adegjk* *bdfhkmn* *ghjloq*
Multipliers by which subsequent blocks are obtained	*ab* *bc* *ac* *cfj* *cfg* *cfh* *cdh*	*bceh* *afk* *bfj* *bek* *aej* *adjm* *efjk*	*acehp* *bcefn* *abfhnp* *cgmpq* *aeghmq* *befgmo* *dfghmnq*

Table 5

No. of factors	10	11	13
Order of replication	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$
No. of plots	128	128	256
Generators of alias subgroup	*ABCDG* *ABEFH* *AGHJK* — —	*ADEKL* *AFGHJ* *BDFJL* *CDGJK* —	*ABCDEF* *DEFGHJ* *GHJKLM* *ADGKN* *BEHKM*
Order of alias of main effects	Third	Third	Third
No. of first-order interactions:			
With third orders as aliases	17	7	34
With second orders as aliases	28	48	44

In the case of the $2^{10/7}$, for confounding in 4 blocks of 32, the confounding subgroup can be

$$I = CEJK = ABGH = ABCEGHJK,$$

and for 8 blocks of 16 the further generator $DFGH$ can be added, when the two first-order interactions AG and CE are lost.

EXPERIMENTS WITH SOME FACTORS AT FOUR LEVELS

A factor at four levels can be regarded as two factors at two levels, the third degree of freedom coming from their interaction. In adopting solutions obtained for factors all at two levels, this interaction must be regarded as a main effect from the point of view of its occurrence in the sets of aliases. This makes the problem of obtaining satisfactory arrangements more difficult, and here only two solutions will be given.

(a) The half-replicate of one factor at four levels and four factors at two levels

We can use $I = ABCDEF$ as the alias subgroup, $I = ABC = ADE = BCDE$ as the confounding subgroup for 4 blocks of 8, and allocate B and D to represent the four-level factor. This loses AF, and leads to the arrangement in Table 6, where the first number in each treatment symbol represents the four-level factors and the remainder the factors A, C, E and F.

Table 6

10000	21011	00111	31100
10011	21000	00100	31111
20111	11100	30000	01011
01101	30110	11010	20001
01110	30101	11001	20010
20100	11111	30011	01000
31010	00001	21101	10110
31001	00010	21110	10101

(b) The half-replicate of two factors at four levels and three factors at two levels

We use the alias subgroup $I = ABCDEFG$, the confounding subgroup

$$I = AEG = CFG = ACEF,$$

and allocate A and B to the first four-level factor and C and D to the second. All the degrees of freedom for the main effects have third-order interactions as aliases, and all the degrees of freedom for the first-order interactions have second-order interactions as aliases. The resulting arrangement is in Table 7.

Table 7

11000	31000	13000	22000
01111	21111	03111	32111
32001	12001	30001	01001
22110	02110	20110	11110
00000	20000	02000	33000
10111	30111	12111	23111
23001	03001	21001	10001
33110	13110	31110	00110
21100	01100	23100	12100
30100	10100	32100	03100
20011	00011	22011	13011
31011	11011	33011	02011
12010	32010	10010	21010
03010	23010	01010	30010
13101	33101	11101	20101
02101	22101	00101	31101

Acknowledgements are due to the Directors of the Distillers Company Limited for permission to publish this paper.

REFERENCES

CARMICHAEL, R. D. (1937). *Introduction to the Theory of Groups of Finite Order*, §28. Boston: Ginn.

FINNEY, D. J. (1945). The fractional replication of factorial arrangements. *Ann. Eugen., Lond.*, **12**, 291–301.

FINNEY, D. J. (1946). Recent developments in the design of field experiments. III. Fractional replication. *J. Agric. Sci.* **36**, 184–91.

FISHER, R. A. (1942*a*). *The Design of Experiments*, 3rd ed., §45.2. London: Oliver and Boyd.

FISHER, R. A. (1942*b*). The theory of confounding in factorial experiments in relation to the theory of groups. *Ann. Eugen., Lond.*, **11**, 341–53.

KEMPTHORNE, O. (1947). A simple approach to confounding and fractional replication in factorial experiments. *Biometrika*, **34**, 255–72.

THE RELATIONSHIP BETWEEN FINITE GROUPS AND COMPLETELY ORTHOGONAL SQUARES, CUBES, AND HYPER-CUBES

BY K. A. BROWNLEE AND P. K. LORAINE

The Research Department of the Distillers Company, Ltd., Epsom, Surrey

Completely orthogonal squares of side n afford a means of introducing $(n+1)$ factors all at n levels into an experiment involving n^2 plots instead of the n^{n+1} that would be required for a full factorial experiment. It is pertinent to inquire how far interactions between pairs of factors will affect the estimates of the main effects of the other factors. The theory of fractional replication, as based on the theory of finite prime power groups by Finney (1945), allows an approach to be made to this problem. Completely orthogonal hyper-cubes will also be investigated.

Finney has indicated the theory of fractional replication of experiments involving n factors each at π levels, where π is any prime. The treatment combinations can be represented by symbols $a^\alpha b^\beta c^\gamma \ldots$, where the indices $\alpha, \beta, \gamma, \ldots$ take only the values $0, 1, 2, \ldots (\pi-1)$. If the symbols are interpreted according to the ordinary laws of algebra, with the additional condition

$$a^\pi = b^\pi = c^\pi = \ldots = 1,$$

then it is found that the product of any two symbols is a third. The complete set of symbols for a full factorial experiment form a prime power group of modulus π and order π^n.

Two symbols $a^\alpha b^\beta c^\gamma \ldots$ and $a^{\alpha'} b^{\beta'} c^{\gamma'} \ldots$ are said to be orthogonal if

$$\alpha\alpha' + \beta\beta' + \gamma\gamma' + \ldots = 0 \pmod{\pi}.$$

If we have a subgroup of order π^p it is possible to select a second subgroup of order π^{n-p} such that all its elements are orthogonal to those of the first subgroup; this second subgroup is said to be the complete orthogonal subgroup of the first subgroup. Finney has shown that if the first subgroup represents the treatments carried out in a fractional replication of a factorial experiment then the effects are interconfused in sets of $(n-p)$, and to find the aliases of any effect we multiply the second subgroup by that effect. The second subgroup is therefore called the alias subgroup.

FORMATION OF COMPLETELY ORTHOGONAL SQUARES OF ANY PRIME NUMBER

To construct a completely orthogonal square of side π, where π is a prime, we need to allocate to each of the π^2 plots their row and column numbers and their levels of the remaining $(\pi-1)$ factors. The square can be considered either to define the levels of $(\pi+1)$ factors or to define the letters of $(\pi-1)$ alphabets in a square.

Consider the treatment subgroup of π^2 elements generated by $x_1 P_1$, $x_2 P_2$, where

$$P_1 = abc \ldots w, \quad P_2 = ab^2c^3 \ldots w^{\pi-1},$$

and w is the $(\pi-1)$th alphabet. The exponents of x_1 and x_2, reduced to modulus π, will run from 0 to $(\pi-1)$ and can therefore be used to define the cells in the square with co-ordinates y_1, y_2. The element corresponding to cell (y_1, y_2) will be $(x_1 P_1)^{y_1} (x_2 P_2)^{y_2}$ and the levels of the $(\pi-1)$ alphabets are the exponents, reduced to modulus π, of the letters a to w in this product.

It is obvious that every level of each alphabet occurs just once in each row and each column. To see that this square is completely orthogonal, we need also to show that each level of any alphabet occurs once and once only with each level of any other alphabet. Suppose that the levels of the rth and sth alphabets ($r \neq s$) are the same for two different cells (p_1, p_2) and (q_1, q_2). Then if their exponents in P_i are ρ_i and σ_i respectively ($i = 1, 2$), we have

$$\rho_1 p_1 + \rho_2 p_2 = \rho_1 q_1 + \rho_2 q_2, \quad \sigma_1 p_1 + \sigma_2 p_2 = \sigma_1 q_1 + \sigma_2 q_2$$

and since $p_1 \neq q_1$, $p_2 \neq q_2$, this leads to

$$\frac{\rho_1}{\rho_2} = \frac{\sigma_1}{\sigma_2}. \tag{1}$$

Here $\rho_i = 1 + (i-1)(r-1)$, $\sigma_i = 1 + (i-1)(s-1)$, $(i = 1, 2)$,

which gives the contradiction $r = s$. Hence the square is completely orthogonal.

The present treatment using prime power groups is essentially that of Stevens (1939). We can proceed to derive the alias subgroup of order $\pi^{\pi-1}$ completely orthogonal to the above treatment subgroup of order π^2. It can be generated by $X_1 X_2 A^{\pi-1}$ and the $(\pi-2)$ other elements of the form $X_1^t A^{\pi-t-1} \Omega_{t+1}$, where $t = 1, 2, \ldots, \pi-2$ and Ω_u is the uth alphabet. For example, in the case of $\pi = 5$ the treatment subgroup generators are $x_1 abcd$, $x_2 ab^2c^3d^4$, and the alias subgroup generators are $X_1 X_2 A^4$, $X_1 A^3 B$, $X_1^2 A^2 C$, $X_1^3 AD$.

Squares of sides 4 and 8

The present treatment using prime-power groups is not immediately adaptable to the case of squares of sides of the form π^n, but we can deal with the case of 2^2 and 2^3 in the following manner.

Fisher & Yates (1943) give as the treatments for the completely orthogonal square of side 4, writing in order columns, rows and the three alphabets,

11111, 12234, 13342, 14423, 21222, etc.

Let us represent each four-level factor by two two-level factors, so that A, B and AB correspond to the 3 degrees of freedom for the first four-level factor, etc. Thus

$$1 = b, \quad 2 = (1), \quad 3 = a \quad \text{and} \quad 4 = ab.$$

The treatment combinations then can be written

bdfhk, *bgjk*, *bcegh*, *bcdefj*, *d*, etc.

This particular set of elements does not form a subgroup, as it does not contain the identity. It can be converted into a subgroup, however, by multiplying through by any of its elements. Selecting d as the multiplier, we get

bfhk, *bdgjh*, *bcdegh*, *bcefg*, (1), etc.

This subgroup can be generated by

bfhk, *bdgjh*, *cehjk*, *aegj*.

We have 5 four-level factors or 10 two-level factors, so the full factorial experiment will have $4^5 = 2^{10}$ elements and the alias subgroup completely orthogonal to the treatment subgroup will have $2^{10-4} = 2^6$ elements and can be generated by

BDF, *ACE*, *BEGH*, *AEFH*, *AFJK*, *DHK*.

The treatment of squares of side 8 is similar to those of side 4. We represent the first 8-level factor by the three 2-level factors A, B and C, etc., such that

$$1 = (1), \quad 2 = a, \quad 3 = b, \quad 4 = ab, \quad 5 = c, \quad 6 = ac, \quad 7 = bc, \quad 8 = abc.$$

The treatment subgroup of $8^2 = 2^6$ elements can be generated by

dgmopqrstuwyα, *ehkprstuvwxzβ*, *fjlmnopqrtvxz*, *agkngtwz*, *bhloruxα*, *cjmpovyβ*.

The complete group corresponding to the full factorial experiment will have $8^9 = 2^{27}$ elements, so the alias subgroup completely orthogonal to the above treatment subgroup will have $2^{27-6} = 2^{21}$ elements.

The selection of 21 elements from the 2,097,152 to act as generators is not too easy, but a systematic method of approach leads to

CDFN, EMP, DNQ, DHV, DTZ, BFL, FGHJKLM,
FPS, BGNO, CMSV, FTN, DRX, CHOP, VFβ,
AHQR, NWXα, FMY, JNPT, ADG, BEH, CFJ.

SQUARES WITH LESS THAN THE COMPLETE NUMBER OF FACTORS

As they represent the most general condition, we have discussed the completely orthogonal squares. The above treatments can, however, be immediately adapted to squares with less than the complete number of factors.

To consider a Latin square of side 5, for example, the complete treatment group for the full factorial experiment has 5^3 elements. The treatment subgroup as before has 5^2 elements, and these can be obtained from those previously given by missing out the last three letters, i.e. the generators are x_1a, x_2a. The alias subgroup will now have $5^{3-2} = 5$ elements and may be generated from $X_1X_2A^4$.

For a Graeco-Latin square the alias subgroup will have $5^{4-2} = 5^2$ elements, and we add the second generator given previously, X_1A^3B. Similarly, for a hyper-Graeco-Latin square the alias subgroup has $5^{5-2} = 5^3$ elements, and we add the third generator given previously, $X_1^2A^2C$.

In all cases above for each factor omitted the appropriate treatment subgroup is obtained by missing out from the generators the letters corresponding to the omitted factor, and the appropriate alias subgroup is obtained by omitting those generators containing letters corresponding to the omitted factor.

RELATIONSHIP BETWEEN MAIN EFFECTS AND INTERACTIONS BETWEEN OTHER FACTORS

In all cases for squares ranging from Latin to completely orthogonal the alias subgroups contain three-letter elements of the type $A^\alpha B^\beta C^\gamma$, and every factor occurs in such elements. Thus all the degrees of freedom for the A main effect have as aliases terms of the type $B^\beta C^\gamma$, i.e. first-order interactions. In using orthogonal squares for experimental designs it is therefore essential that no pair of factors should be liable to interact. In view of the unsatisfactory nature of this result, one is led to inquire whether Latin or completely orthogonal cubes or hyper-cubes will provide a sounder basis for experimental designs.

ORTHOGONALITY IN THREE DIMENSIONS

Consider a treatment subgroup generated by elements of the form x_1P_1, x_2P_2, x_3P_3, where

$$P_1 = abcd \dots w, \quad P_2 = ab^2c^3d^4 \dots w^{\pi-1}, \quad P_3 = ab^3c^5d^7 \dots w^{2\pi-3},$$

in which w is the $(\pi-1)$th letter, and whenever the exponent of one of the letters a to w becomes 0 (mod π) such a letter is deleted from all the generators. The letter to be deleted will in fact be the $\left(\frac{\pi+1}{2}\right)$th. This treatment subgroup of π^3 elements can be considered to define the levels of $(\pi+1)$ factors or the levels of $(\pi-2)$ alphabets in a cube, the position

co-ordinates in this cube being given by the x_i. The alphabets are numbered before the deletion is made and for convenience this numbering will be retained. The levels of the alphabets in cell (y_1, y_2, y_3) are given by the exponents of the letters a to w in the element $(x_1P_1)^{y_1}(x_2P_2)^{y_2}(x_3P_3)^{y_3}$.

The condition for such a cube to be Latin is that in any 'file' defined by two of the y_i being constant, each of the π levels of each alphabet should occur once and once only. This condition is obviously fulfilled here.

For the cube to be completely orthogonal we must also have that each plane section $y_i = \text{const.}$ $(i = 1, 2, 3)$ forms an orthogonal square. The arguments used above to obtain (1) hold here if suffices j and k are substituted for 1 and 2 throughout $(j, k = 1, 2, 3)$. Therefore we have

$$\frac{\rho_j}{\rho_k} = \frac{\sigma_j}{\sigma_k},$$

where $\rho_i = 1 + (i-1)(r-1)$, $\sigma_i = 1 + (i-1)(s-1)$, $(i = j, k)$, which leads to $r = s$ as before.

The alias subgroup will be of order $\pi^{\pi-2}$ and has $(\pi-2)$ generators of the form

$$X_1X_2X_3A^{\pi-1}, \quad X_1^t X_3^{\pi-t} A^{\pi-t-1}\Omega_{t+1},$$

where $t = 1, 2, \ldots, \dfrac{\pi-1}{2}, \dfrac{\pi+3}{2}, \ldots, \pi-2$, and Ω_u is the uth alphabet.

Orthogonality in four or more dimensions

Considerations analogous to those for three dimensions show that for four dimensions the treatment subgroup generators will be

$$x_1abc\ldots w, \quad x_2ab^2c^3\ldots w^{\pi-1}, \quad x_3ab^3c^5\ldots w^{2\pi-3}, \quad x_4ab^4c^7\ldots w^{3\pi-5},$$

where, as before, whenever the exponent of one of the letters a to w becomes 0 (mod π), such a letter is deleted from all generators. Thus the $\left(\dfrac{\pi+1}{2}\right)$th letter will be deleted and also the $\left(\dfrac{\pi+2}{3}\right)$th or the $\left(\dfrac{2\pi+2}{3}\right)$th, according as π is of the form $(3k+1)$ or $(3k+2)$, k being an integer.

The generalization to n dimensions is now obvious. The treatment subgroup will have n generators of the form x_iP_i, where

$$P_i = ab^{1+(i-1)}c^{1+2(i-1)}\ldots w^{1+(\pi-2)(i-1)} \quad (i = 1, 2, \ldots, n),$$

deletions being made as before. After deletions there will thus always be a total of $(\pi+1)$ classifications or factors.

The proofs of complete orthogonality follow exactly as for three dimensions.

The treatment subgroup is of order π^n and therefore the alias subgroup is of order $\pi^{\pi+1-n}$ and has generators of the form

$$X_1X_2\ldots X_nA^{\pi-1},$$
$$X_1^{\pi+t}X_2^{\pi+0.t}X_3^{\pi-1.t}X_4^{\pi-2.t}\ldots X_n^{\pi-(n-2)t}A^{\pi-t-1}\Omega_{t+1},$$

where Ω_u is the uth alphabet and, if $m_1, m_2, \ldots, m_{n-2}$ are the numbers in ascending order of the omitted alphabets, then

$$t = 1, 2, \ldots, (m_1-1), (m_1+1), \ldots, (m_2-1), (m_2+1), \ldots, (m_{n-2}-1), (m_{n-2}+1), \ldots, \pi-2.$$

EXPERIMENTAL DESIGNS: LATIN HYPER-CUBES

Inspection of the generators for the treatment subgroups shows that a Latin (i.e. having only one alphabet) hyper-cube in any number of dimensions is always possible for any prime. For example, we can have a Latin hyper-cube in four dimensions for $\pi = 3$. This will have 5 classifications and be based on 3^4 cells; it is thus a solution of the one-third replicate of 5 factors at three levels. This solution is not actually the same as given by Finney (1945) as he selected one more satisfactory for confounding. Similarly the one-fifth replicate of 6 factors at five levels can be generated immediately and its alias subgroup will have the one generator $X_1X_2X_3X_4X_5A^4$ thus allowing first-order interactions to be estimated with third orders as aliases.

EXPERIMENTAL DESIGNS: GRAECO-LATIN HYPER-CUBES

In view of the form of the generators for the treatment subgroup, and in particular the property of elimination of letters where exponents become 0 (mod π), it is clearly impossible to generate, with this method, n-dimensional hyper-cubes other than Latin except when $\pi \geqslant n+1$.

In the case of Graeco-Latin hyper-cubes we have two alphabets and n co-ordinates making up $(2+n)$ factors in π^n plots. The order of replication is therefore π^{-2}, and the alias subgroup, generated by

$$X_1X_2 \dots X_nA^{\pi-1}, \quad X_1X_3^{\pi-1} \dots X_n^{\pi-(n-2)}A^{\pi-2}B,$$

will have no element with less than $(n+1)$ letters.

In the case of three dimensions, for $\pi = 3$ only a Latin cube is possible. With $\pi \geqslant 5$, the Graeco-Latin is possible, and the main effects of all factors will have second-order interactions as aliases. For example, with $\pi = 5$ we can have 5 factors with $5^3 = 125$ plots.

In the case of four dimensions, π must be $\geqslant 5$. For example, with $\pi = 5$, the two alphabets and four co-ordinates make 6 factors in $5^4 = 625$ plots. Main effects have third-order interactions and first-order interactions have second-order interactions as their lowest order aliases.

On going to five dimensions the minimum prime is 7 and hence the minimum number of plots is $7^5 = 16807$.

EXPERIMENTAL DESIGNS: COMPLETELY ORTHOGONAL HYPER-CUBES

Consideration of the general alias subgroup shows that if this is of order π^m, $m \geqslant 3$, then its generators must include $(m-1)$ of the type

$$X_1^t X_3^{\pi-t} \dots X_n^{\pi-(n-2)t}A^{\pi-t-1}\Omega_{t+1}.$$

Since the subgroup generated by any pair of elements of this type contains elements of the form $A^\alpha B^\beta C^\gamma$, and since every alphabet occurs in such elements, all the factors corresponding to the alphabets include first-order interactions amongst their aliases. The factors corresponding to the co-ordinates will be better, however, having as aliases interactions of order $(n-1)$.

For example, with $\pi = 5$, in three dimensions, we can have three alphabets. The total number of plots is $5^3 = 125$ and the order of replication $5^{-3} = 1/125$. Three factors have first-order aliases and three factors have second-order aliases. With $\pi = 7$, the total number of plots is $7^3 = 343$, being a $7^{-5} = 1/16807$ replicate. Five factors have first-order interactions as aliases and three have second-order aliases.

ADDITIONAL NOTE ON 'THE DESIGN OF OPTIMUM MULTIFACTORIAL EXPERIMENTS'

Kempthorne (1947) has pointed out that the multifactorial designs of Plackett & Burman (1946) are in some cases high-order fractionally replicated factorial designs.

Thus their case of $N = 16$, $L = 2$, $n = 15$ is a 2^{-11} replicate of the 2^{15} experiment. The alias subgroup of 2^{11} elements for their design can be generated by

$$ABN,\ ACK,\ ADE,\ AFL,\ AGJ,\ AHO,\ AMP,\ BCO,\ BDL,\ BGM,\ CDP.$$

Altogether there are 35 three-letter elements in the alias subgroup, and each factor has, amongst its aliases, 7 first-order interactions involving the other 14 factors.

Their case of $N = 32$, $L = 2$, $n = 31$ has an alias subgroup of $2^{31-5} = 2^{26}$ elements containing 155 three-letter elements. Thus each factor has, amongst its aliases, 15 first-order interactions involving the other 30 factors.

The three cases where $N = 9$, 25 and 49, $L = 3$, 5 and 7, and $n = 4$, 6 and 8 respectively are completely orthogonal squares with the alias characteristics described above.

With $N = 27$, $L = 3$, $n = 13$, if we ascribe the letters a to n (omitting i) to the 13 factors in the order given, the treatments are a subgroup generated by

$$ad^2f^2gh^2j^2km^2n^2,\quad bde^2fgh^2klm,\quad cef^2ghj^2lmn,$$

and the alias subgroup of 3^{10} elements is generated by

$$BCL^2,\ A^2B^2K,\ D^2EG^2,\ EGN,\ CEL,\ C^2K^2G,\ B^2J^2F,\ ACJ,\ AL^2M,\ C^2HK.$$

Thus in all the multifactorial designs amenable to representation as fractional replications, all main effects contain first-order interactions amongst their aliases.

Thanks are due to the Distillers Company Limited for permission to publish this paper.

REFERENCES

FINNEY, D. J. (1945). The fractional replication of factorial arrangements. *Ann. Eugen., Lond.*, **12**, 291–301.

FISHER, R. A. & YATES, F. (1948). *Statistical Tables for Biological, Agricultural, and Medical Research* (3rd edition). London: Oliver and Boyd.

KEMPTHORNE, O. (1947). A simple approach to confounding and fractional replication in factorial experiments. *Biometrika*, **34**, 255–72.

PLACKETT, R. L. & BURMAN, J. P. (1946). The design of optimum multifactorial experiments. *Biometrika*, **33**, 305–25.

STEVENS, W. L. (1939). The completely orthogonalized Latin square. *Ann. Eugen., Lond.*, **9**, 82–93.

SYSTEMATIC SAMPLING OF CONTINUOUS PARAMETER POPULATIONS

A. E. JONES

[*Editorial note.* This paper was submitted by Dr Jones in the autumn of 1947. A number of points arising on the original draft were discussed with him, and the Editors were about to consult him on further amendments which they thought desirable when Dr Jones met an untimely death in a lift accident on 7 May 1948. The amendments have, nevertheless, been made in the interests of clarity, and the more important of them are indicated in this paper. No alteration has been made in the substance of the paper or in the main formulae, but the second part has been omitted and is discussed in the succeeding paper by Mr Kendall.]

Introduction

1. A sample is usually said to be distributed systematically when the individual members are selected according to some deterministic rule. For instance, if it were desired to estimate the mean of a characteristic of a mass-produced article, and a sample consisting of every tenth article were taken, we might say that this was a *systematic* one. Such a sample may, of course, be random if the method of selection of the samples is unconnected, in any conceivable manner, with the value of the characteristics under consideration.

Estimates obtained from a systematic sample may sometimes be more accurate than those from a random sample of the same size, but it is usually considered that this lower accuracy is more than counterbalanced by the fact that a more reliable estimate of the sampling variation can be obtained by the latter method. The customary formulae giving an estimate of sampling error are based on the assumption that there is no correlation between the individual members of the sample. It sometimes happens that neighbouring sample members are positively correlated, in which case the error estimated from the customary formulae may differ substantially from the true value.

The problem to be considered here is that of estimating by a systematic sample the 'average value' of a random variable from a one-dimensional homogeneous population depending on a continuous parameter, such as the nitrogen content of soil along a one-dimensional strip. The phrase 'average value' requires closer definition. Suppose $Z(t)$ be used to denote the actual nitrogen content at a fixed point distance t from one end of the strip under consideration. Then $Z(t)$ may be regarded as a random variable with a hypothetical population consisting of all the nitrogen contents at this point throughout a number of years. There is an infinity of random variables corresponding to all the points of the t-interval considered. We are not interested in the means of these populations, but in the value of the nitrogen content of the whole of the strip of soil. Thus, if T_0 and T $(> T_0)$ are the ends of the interval, the average value in which we are interested is

$$\bar{Z} = \frac{1}{T - T_0} \int_{T_0}^{T} Z(t)\, dt,$$

which is not a parameter of any of the hypothetical populations mentioned above, and, in fact, is itself a random variable.

It will only be possible to make a finite number of observations and if, say, n is the maximum number it is proposed to make, we would wish to dispose them in such a way that

the mean-square error of our estimate of $\bar{Z}$ will be a minimum. Hence if $z(t_1), z(t_2), \ldots, z(t_n)$ represent observed variates at $t_1, t_2, \ldots, t_n$ respectively, and if our best* estimate of $\bar{Z}$ is the linear combination

$$a_1 z(t_1) + a_2 z(t_2) + \ldots + a_n z(t_n),$$

the mean-square error of the estimate will be

$$S = E[a_1 z(t_1) + a_2 z(t_2) + \ldots + a_n z(t_n) - \bar{Z}]^2, \tag{1·1}$$

where the operator E denotes expected value.

Apart from determining $t_1, t_2, \ldots, t_n$ we also require the mean-square error of this estimate of $\bar{Z}$. Mathematically, this resolves into two problems: first, that of obtaining the mean-square error as a function of the (unknown) parameters of the distribution of $Z(t)$, and secondly obtaining a formula in terms of actual sample values which will provide a satisfactory estimate of this function.

It will be shown, on reasonable hypotheses, that if the correlation between the successive values at the points $t_1, t_2, \ldots, t_n$ is very small, the best method of distributing n members along a strip of length T is at distances $T/(n+1)$, $2T/(n+1)$, ... from one end, while if the correlation is greater than 0·25 the best distribution will be at distances $T/2n$, $3T/2n$, ...

Theoretical representation of $Z(t)$—assumptions

2. It is necessary to make certain assumptions about the way in which the random function $Z(t)$ varies. Physically, we should expect that there will be some smooth curve, or trend, about which $Z(t)$ will fluctuate randomly. In addition, we should expect that $Z(t)$ will be continuous in t, in other words we should expect that by making h small we could make $[Z(t+h) - Z(t)]$ as small as we please. Incidentally, it follows from this that it is impossible to regard the random variables $Z(t)$ and $Z(t+h)$ as independent for all values of h.

Thus any recorded value, $z(t_1)$, is the sum of three components: (*a*) an error of observation; (*b*) the value of the trend at t_1; (*c*) a random variable, which will be denoted by $X(t_1)$, representing $Z(t_1)$ less the value of the trend at t_1. From the definition of the trend we may take

$$EX(t_1) = 0 \quad \text{for all } t_1. \tag{2·1}$$

The following assumptions will be made about these three components:

(A) That the error of observation, i.e. $z(t_1) - Z(t_1)$, is a random variable with zero mean, and is uncorrelated with all other errors of observation, and with $Z(t)$ for all t. $x(t_1)$ will be used to denote the sum of $X(t_1)$ and the error of observation at t_1. In other words, $x(t_1)$ is $z(t_1)$ corrected for trend.

(B) In the first place it will be assumed that the trend is simply a constant for all t. If

$$\bar{X} = \frac{1}{T - T_0}\int_{T_0}^{T} X(t)\,dt, \tag{2·2}$$

the problem of minimizing S of equation (1·1) then reduces to the problem of minimizing

$$E\{a_1 x(t_1) + a_2 x(t_2) + \ldots + a_n x(t_n) - \bar{X}\}^2, \tag{2·3}$$

* Throughout this paper the expressions 'best estimate' or 'optimum estimate' will be used to mean the linear estimate whose mean-square error is a minimum. [The expression 'mean-square error' is used instead of 'variance' because we are considering a minimization of the type $E(X-Y)^2$, where X and Y are random variables, not of the type $E\{X - E(x)\}^2$.—Ed.]

provided we confine our attention to unbiased estimates, i.e. make the restriction*

$$\sum_{j=1}^{n} a_j = 1. \tag{2·4}$$

(C) About $X(t)$ it will be assumed that:

(1) The variance of $X(t) = E\{X(t)\}^2 = \lambda$, a constant independent of t. (2·5)

(2) The correlation of $X(s)$ and $X(t)$ depends only on $|s-t|$, and decreases exponentially as $|s-t|$ increases. $E\{X(s)\,X(s+t)\}$ will be denoted by $\lambda R(t)$, where $R(t)$ is the correlation function of $X(t)$. Thus

$$R(t) = e^{-q|t|} \quad (q>0). \tag{2·6}$$

It follows that for h small

$$E\{X(t+h)-X(t)\}^2 = 2\lambda(1-e^{-q|h|}) \tag{2·7}$$

is also small, and since

$$E\{X(t+h)-X(t)\} = 0,$$

it will be seen that by these assumptions we imply that $X(t)$ is continuous in the mean for all t.†

I consider that the above assumptions are the simplest that satisfy the required physical conditions. (A) represents nothing new. Although it is assumed that the trend is a constant for all t, the results are valid if the trend is a linear function of t. As regards the assumption (C) it is worth noting, though it will not be discussed here, that there is a 'small-error' justification for (2·5) and (2·6) similar to the usual small-error justification for the assumption of normality. They are both derived from the Central Limit Theorem.

I have made some practical attempts to decide whether the above assumptions are satisfactory for a number of meteorological variables, e.g. wind speed and air temperature. It was found that the observations analysed could be satisfactorily interpreted on these hypotheses, but a really thorough examination of their appropriateness would require computation far beyond the author's resources.‡

SYSTEMATIC SELECTION OF SAMPLING POINTS

3. Suppose that sample members are selected at points $t_1, t_2, \ldots, t_n$ in that order, and let $X_i = X(t_i)$ be the observed variables corrected for trend. Also let x_i denote the corresponding observations corrected for trend. Then the expected value of $(x_i - X_i)$ will be zero.

Let

$$E(x_i - X_i)^2 = \mu.$$

Then, expressed mathematically, the problem with which we shall be concerned here is that of estimating the random variable, I, where

$$I = (T-T_0)\,\bar{X} = \int_{T_0}^{T} X(t)\,dt,$$

from the random variables, $x_1, x_2, \ldots, x_n$.

* [Since $Ex(t) = 0$ the expected value of $\Sigma a_j x(t_j)$ is zero, not $\bar{X}$. The author means that the average of $\Sigma a_j x(t_j)$ *over all* t is equal to $\bar{X}$ which gives his condition that $\Sigma a_j = 1$.—ED.]

† If, in addition, we make the assumption (which is irrelevant to the subsequent development of this paper) that $X(t)$ is Gaussian, then $X(t)$ must be a stationary Gaussian-Markoff random function. It has been shown that such functions are everywhere continuous.

‡ [It was the Editor's intention to suggest that Dr Jones should include some of this experimental verification in his paper. The author assumes that $X(t)$ has the same mean, variance and correlation function for all t. His problem is analogous to the estimation of the stationary stochastic component of a continuous time-series defined at all points of time as the sum of three components, a linear trend, a random variable which possesses a correlogram decaying exponentially to zero and a 'superposed' error.—ED.]

Now since by 2 (A) and 2 (C)

$$E[(x_i - X_i)(x_j - X_j)] = 0 \quad (i \neq j), \tag{3·1}$$

$$E[X(s+t)\,X(s)] = \lambda e^{-q|t|}, \tag{3·2}$$

the correlation matrix of the vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is given by

$$\left.\begin{aligned} E(x_i x_j) &= \lambda e^{-q|t_i - t_j|} \quad (i \neq j) \\ &= \lambda + \mu \quad (i = j). \end{aligned}\right\} \tag{3·3}$$

If $$P_i = e^{-qt_i}, \quad P_0 = e^{-qT_0}, \quad P_{n+1} = e^{-qT},$$

then $$E(\mathbf{x}\mathbf{x}') = \lambda R + \mu I_n, \tag{3·4}$$

where $$R = \begin{pmatrix} 1 & P_2/P_1 & P_3/P_1 & P_n/P_1 \\ P_2/P_1 & 1 & P_3/P_2 & P_n/P_2 \\ P_3/P_1 & P_3/P_2 & 1 & P_n/P_3 \\ \ldots & \ldots & \ldots & \ldots \\ P_n/P_1 & P_n/P_2 & \ldots & 1 \end{pmatrix}$$

and I_n is the unit matrix.*

Also, denoting $E(x_i I)$ by n_i, we have that

$$\begin{aligned} n_i &= E\left\{\int_{T_0}^{T} X(t)\,x_i\,dt\right\} \\ &= \int_{T_0}^{T} E[X(t)\,x_i]\,dt \\ &= \frac{\lambda}{q}\left(2 - \frac{P_i}{P_0} - \frac{P_{n+1}}{P_i}\right). \end{aligned} \tag{3·5}$$

THEOREM 1. Subject to the restriction that the sum of the weights, v_i, is $(T - T_0)$, the best estimate of I that can be obtained from any n samples is $\sum_{i=1}^{n} v_i x_i$, the sample values x_i being selected at points t_i $(t_i > t_{i-1})$, where

(*a*) $$v_1 = v_2 = \ldots = v_n = \frac{T - T_0}{n}. \tag{3·6}$$

(*b*) $$\begin{aligned} t_2 - t_1 = t_3 - t_2 &= \ldots = t_n - t_{n-1} \\ &= \frac{1}{q}\log_e\left\{\left(k + \frac{\lambda - \mu}{2\lambda}\frac{T - T_0}{n}q\right)\Big/\left(k - \frac{\lambda - \mu}{2\lambda}\frac{T - T_0}{n}q\right)\right\}. \end{aligned}$$

(*c*) $$t_1 - T_0 = T - t_n = -\frac{1}{q}\log_e\left(k - \frac{\lambda + \mu}{2\lambda}\frac{T - T_0}{n}q\right), \tag{3·7}$$

and k satisfies the equation

$$\left(k - \frac{\lambda + \mu}{2\lambda}\frac{T - T_0}{n}q\right)^{n+1}\Big/\left(k + \frac{\lambda - \mu}{2\lambda}\frac{T - T_0}{n}q\right)^{n-1} = e^{-q(T - T_0)}. \tag{3·8}$$

Proof. The mean-square error of $\mathbf{v}'\mathbf{x}$ as an estimate of I will be

$$\begin{aligned} S &= E(I - \mathbf{v}'.\,\mathbf{x})^2 \\ &= \mathbf{v}'(\lambda R + \mu I_n)\,\mathbf{v} - 2\mathbf{v}'.\,\mathbf{n} + E(I^2). \end{aligned} \tag{3·9}$$

* The matrix notation used in this paper will be that of *The Theory of Canonical Matrices* by Turnbull and Aitken.

In order that $\mathbf{v}'\mathbf{x}$ be the best estimate, S must be a minimum for t_i and all possible v_i. Considering first the conditions that S should be a minimum by differentiating (3·9) partially with respect to t_i, we obtain by (3·3) and (3·5)

$$e^{-q(t_i-T_0)}-e^{-q(T-t_i)}=-\sum_{j=1}^{i-1}q\,e^{-q(t_i-t_j)}v_j+\sum_{j=i+1}^{n}q\,e^{-q(t_j-t_i)}v_j, \tag{3·10}$$

i.e.
$$\frac{P_i}{P_0}-\frac{P_{n+1}}{P_i}=-q\left\{\sum_{j=1}^{i-1}v_j\frac{P_i}{P_j}+\sum_{j=i+1}^{n}v_j\frac{P_j}{P_i}\right\}\quad(i=1,\ldots,n).$$

These n equations can be combined in the matrix form

$$\frac{P_i}{P_0}-\frac{P_{n+1}}{P_i}=[q(Q-Q')\mathbf{v}]_i\quad(i=1,\ldots,n), \tag{3·11}$$

where
$$Q=\begin{pmatrix}\frac{\lambda+\mu}{2\lambda} & \frac{P_2}{P_1} & \frac{P_3}{P_1} & \ldots & \frac{P_n}{P_1}\\ 0 & \frac{\lambda+\mu}{2\lambda} & \frac{P_3}{P_2} & \ldots & \frac{P_n}{P_2}\\ 0 & 0 & \frac{\lambda+\mu}{2\lambda} & \ldots & \frac{P_n}{P_3}\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ 0 & 0 & \ldots & \ldots & \frac{\lambda+\mu}{2\lambda}\end{pmatrix}.$$

Considering now possible variations of $\mathbf{v}$, the optimum values must satisfy

$$(\lambda R+\mu I_n)\mathbf{v}+k'.\mathbf{i}=\mathbf{n}, \tag{3·12}$$

i.e.
$$2k-\frac{P_{n+1}}{P_i}-\frac{P_i}{P_0}=[q(Q+Q')\mathbf{v}]_i\quad(i=1,\ldots,n), \tag{3·12′}$$

where $\mathbf{i}'=(1,1,\ldots,1)$ and k', k are scalar constants.

The optimum estimate of I which can be made from n samples must thus satisfy (3·11) and (3·12′). Hence

$$k-\frac{P_{n+1}}{P_i}=[qQ\mathbf{v}]_i, \tag{3·13}$$

$$k-\frac{P_i}{P_0}=[qQ'\mathbf{v}]_i\quad(i=1,\ldots,n). \tag{3·14}$$

From (3·13)
$$k(P_i-P_{i+1})=q\left(\frac{\lambda+\mu}{2\lambda}P_iv_i+\frac{\lambda-\mu}{2\lambda}P_{i+1}v_{i+1}\right)\quad(i=1,\ldots,\overline{n-1}). \tag{3·15}$$

Similarly, from (3·14)
$$k(P_i-P_{i+1})=q\left(\frac{\lambda-\mu}{2\lambda}P_{i+1}v_i+\frac{\lambda+\mu}{2\lambda}P_iv_{i+1}\right). \tag{3·16}$$

Combining (3·15) and (3·16)

$$(v_i-v_{i+1})\left(P_{i+1}\frac{\lambda-\mu}{2\lambda}-P_i\frac{\lambda+\mu}{2\lambda}\right)=0,$$

i.e.
$$v_i=v_{i+1}\quad(i=1,2,\ldots,\overline{n-1}). \tag{3·6′}$$

Writing $v_1=v_2=\ldots=v_n=(T-T_0)/n$ in the equations, we have

$$\frac{P_1}{P_2}=\frac{P_2}{P_3}=\ldots=\frac{P_{n-1}}{P_n}=\left(k+\frac{\lambda-\mu}{2\lambda}\frac{T-T_0}{n}q\right)\Big/\left(k-\frac{\lambda+\mu}{2\lambda}\frac{T-T_0}{n}q\right). \tag{3·17}$$

Also we have
$$\frac{P_0}{P_1} = \frac{P_n}{P_{n+1}} = \left(k - \frac{\lambda+\mu}{2\lambda}\frac{T-T_0}{n}q\right)^{-1}. \tag{3·18}$$

Equations (3·17) and (3·18) lead immediately to results (3·7).

Finally, since
$$\frac{P_0}{P_{n+1}} = \frac{P_0}{P_1}\frac{P_1}{P_2}\cdots\frac{P_n}{P_{n+1}} = e^{q(T-T_0)},$$
k must be given by
$$\left(k + \frac{\lambda-\mu}{2\lambda}\frac{T-T_0}{n}q\right)^{n-1} \Big/ \left(k - \frac{\lambda+\mu}{2\lambda}\frac{T-T_0}{n}q\right)^{n+1} = e^{q(T-T_0)}. \tag{3·8}$$

Corollary. It may sometimes happen because the trend is known theoretically, or through some other circumstance, that it is unnecessary to impose the restriction
$$v_1 + v_2 + \ldots + v_n = T - T_0.$$
In this case we obtain by putting $k = 0$, $k = 1$ in the equations (3·13) and (3·14) that the optimum estimate of I is given by
$$v_1 = v_2 = \ldots = v_n = v \text{ (say)},$$
$$\frac{P_1}{P_2} = \frac{P_2}{P_3} = \ldots = \frac{P_{n+1}}{P_n} = 1 + \frac{\lambda-\mu}{2\lambda}vq \Big/ 1 - \frac{\lambda+\mu}{2\lambda}vq,$$
$$\frac{P_0}{P_1} = \frac{P_n}{P_{n+1}} = \left(1 - \frac{\lambda+\mu}{2\lambda}vq\right)^{-1},$$
$$\left(1 + \frac{\lambda-\mu}{2\lambda}vq\right)^{n-1} \Big/ \left(1 - \frac{\lambda+\mu}{2\lambda}vq\right)^{n+1} = e^{q(T-T_0)}.$$

Distribution of sample values—practical application

4. Two simple conclusions emerge from equations (3·6)–(3·8). They are that the optimum estimate of the 'average value' for a given sample size is one in which the same sample values are combined with equal weight $1/n$ and in which the distances between successive sample values are all equal, i.e. $t_2 - t_1 = t_3 - t_2 \ldots$. The distance of the first sample value from the end, $t_1 - T_0$, is governed by a rather more complicated equation, but it will be shown that $t_1 - T_0$ must lie between $\frac{1}{2}(t_2 - t_1)$ and $(t_2 - t_1)$. It is worth while to note, in passing, that though the formulae for $t_1, t_2, \ldots, t_n$ have been worked out on the assumption that the trend (§ 2) is simply a constant, *the formulae will still be accurate if the trend is a linear function of* t.* In fact no other distribution of sample values is likely to decrease appreciably the error in the estimate of mean arising from the possible curvature of the trend.

It is only necessary to know the ratio $(t_1 - T_0)/(t_2 - t_1)$ to be able to specify the positions $t_1, t_2, \ldots, t_n$ completely.

Theorem 2. If the distribution of sample values which gives the best estimate of the mean is at points $t_1, t_2, \ldots, t_n$ $(t_i > t_{i-1})$, then
$$t_2 - t_1 = t_{i+1} - t_i \quad (i = 2, 3, \ldots, n-1), \qquad t_1 - T_0 = T - t_n$$

* Suppose the trend could be expressed functionally as $A + Bt$. Then the truth of the above statement can readily be seen by considering the random function $X(t) = Z(t) - A - Bt$, and applying the formula,
$$E\left[\sum_{i=1}^{n} a_i z_i - \int_{T_0}^{T} Z(t)\,dt\right]^2 = E\left[\sum_{i=1}^{n} a_i x_i - \int_{T_0}^{T} X(t)\,dt\right]^2 + \left[\sum_{i=1}^{n} a_i t_i - \int_{T_0}^{T} t\,dt\right]^2.$$

and (*a*) $(t_1-T_0)/(t_2-t_1)$ is independent of λ and μ,

(*b*) $\frac{1}{2}(t_2-t_1) \leqslant t_1-T_0 \leqslant t_2-t_1$, (4·1)

(*c*) $(t_1-T_0)/(t_2-t_1)$ can be determined very closely from the correlation of actual values, distance $(T-T_0)/n$ apart, i.e. from

$$e^{-q(T-T_0)/n}, \tag{4·2}$$

and is otherwise approximately independent of n.

Proof of (*a*). If $(T-T_0)\,q/n = r \qquad k-\frac{1}{2}\mu r/\lambda = K,$ (4·3)

$$p = e^{-q(t_i-t_{i-1})} \quad (i = 2, \ldots, n), \qquad p_0 = e^{-q(t_1-T_0)}. \tag{4·4}$$

From (3·15)–(3·18) $p_0 = K-\frac{1}{2}r,$ (4·5)

$$p = (K-\tfrac{1}{2}r)/(K+\tfrac{1}{2}r), \tag{4·6}$$

$$p_0^2 p^{n-1} = (K-\tfrac{1}{2}r)^{n+1}/(K+\tfrac{1}{2}r)^{n-1} = e^{-nr}. \tag{4·7}$$

The value of K is given by (4·7) and $(t_1-T_0)/(t_2-t_1)$ can be determined from (4·5) and (4·6). They are clearly independent of λ and μ.

Proof of (*b*). We have, from (4·6),

$$K-\tfrac{1}{2}r = rp/(1-p) = p_0. \tag{4·5′}$$

Hence by (4·7) $$p^{n+1}/(1-p)^2 = e^{-nr}/r^2. \tag{4·8}$$

Now $$p^{-1}-p^1 \geqslant 2\log_e 1/\sqrt{p} = \log_e 1/\sqrt{p}.$$

Hence $$\frac{p}{(1-p)^2} \leqslant \frac{1}{(\log_e 1/p)^2}$$

$$\frac{e^{-nr}}{r^2} = \frac{p^{n+1}}{(1-p)^2} \leqslant \frac{e^{-n\log_e 1/p}}{(\log_e 1/p)^2}.$$

Now e^{-nr}/r^2 is a monotone decreasing function of r $(r>0)$. Hence $\log_e 1/p \leqslant r$, and since $p_0^2 p^{n-1} = e^{-nr}$, it follows that $\log_e 1/p_0 \geqslant \frac{1}{2}r$. Hence

$$\frac{t_1-T_0}{t_2-t_1} \geqslant \tfrac{1}{2}. \tag{4·1′}$$

Also, since $1-e^{-rn/(n+1)} \leqslant r$, by (4·7),

$$p(1-p^2)^{-2/(n+1)} = r^{-2/(n+1)}\,e^{-rn/(n+1)} \leqslant \frac{e^{-rn/(n+1)}}{(1-e^{-rn/(n+1)})^{2/(n+1)}}.$$

Now $p(1-p^2)^{-2/(n+1)}$ is a monotone increasing function of p $(0<p<1)$. Hence $p \leqslant e^{-rn/(n+1)}$. But from (4·7) $p_0 \geqslant e^{-rn/(n+1)}$. So by (4·4)

$$\frac{t_1-T_0}{t_2-t_1} \leqslant 1. \tag{4·1″}$$

Thus (4·1′) and (4·1″) establish (4·1).

Proof of (*c*). To prove this result, we show that for a given r, $(t_1-T_0)/(t_2-t_1)$ is a monotone function of n. For

$$\frac{t_1-T_0}{t_2-t_1} = \frac{-\log(K-\frac{1}{2}r)}{-\log(K-\frac{1}{2}r)+\log(K+\frac{1}{2}r)}. \tag{4·9}$$

It follows from (4·2) that $K+\frac{1}{2}r>1$. Also K is a monotone function of n. Hence, when n increases $(t_1-T_0)/(t_2-t_1)$ decreases.

Some values of this ratio have been worked out and are presented in Table 1.* For $n=3$ the value is nearly the same as for n large and thus the ratio is nearly independent of n. This completes the proof of theorem 2.

Table 1 (*see text*)

$T-T_0$ = length of the interval sampled; n = number of sample members. L is the distance such that the correlation of true values, distance L apart, will be 0·5, i.e. is given by $e^{-qL}=0{\cdot}5$.

$\frac{T-T_0}{nL}$	p = correlation between successive sample members	$\frac{t_1-T_0}{t_2-t_1}$ (n large)	$\frac{t_1-T_0}{t_2-t_1}$ ($n=3$)
0·5	0·707	0·51	0·51
1·0	0·500	0·53	0·51
2·0	0·250	0·56	0·54
3·0	0·125	0·58	0·56
4·0	0·062	0·61	0·58
5·0	0·031	0·63	0·60
7·0	0·008	0·67	0·65
10·0	0·001	0·72	0·69
15·0	$0{\cdot}0^43$	0·77	0·74
20·0	$0{\cdot}0^51$	0·82	0·78

We may expand the ratio $(t_1-T_0)/(t_2-t_1)$ in terms of r. It will be sufficient to consider the expansion when n is large as, by theorem 2, the result is nearly independent of n. We have

$$\frac{t_1-T_0}{t_2-t_1}=\frac{\log\{pr/(1-p)\}}{\log p}=\frac{r+\log\{r/(1-p)\}}{r}$$

$$=\frac{1}{2}+\frac{r}{24}+O(r^3). \tag{4·10}$$

A similar expansion in terms of the parameter $1-e^{-r}$, which is nearly equal to $1-p$, is

$$\frac{1}{2}+\frac{1-e^{-r}}{24}+\frac{(1-e^{-r})^2}{48}+O(1-e^{-r})^3. \tag{4·11}$$

We cannot expect to have anything but a rough idea of r before an experiment is conducted. Fortunately, this is all that will be required. It will be seen from Table 1 that if $e^{-r}>0{\cdot}25$ ($r>2{\cdot}8$), no accuracy will be lost if the sample members are placed at distances $(T-T_0)/2n$, $3(T-T_0)/2n$, etc., from T_0. In other words, if the correlation between successive sample members is likely to be appreciable, we divide the whole interval into n equal sections, and place one member at each of the mid-points of the sections. Again, if the number of sample members is large, then the loss of accuracy resulting from placing the samples at distances $(T-T_0)/2n$, $3(T-T_0)/2n$ from T_0 should be small.

* [*Editorial note.* Dr Jones appears to have derived the last column of Table 1, first by determining r by successive approximation from p according to the relation $e^{-nr}/r^2=p^{n+1}/(1-p)^2$, and then by the use of

$$(t_1-T_0)/(t_2-t_1)=\log\{pr/(1-p)\}/\log p.$$

The approximation of (4·10) is not very useful for small p because r is then large. It may be shown that for large n the ratio $(t_1-T_0)/(t_2-t_1)$ tends to $1-\{\log(1-p)-\log|\log p|\}/\log p$ which obviates the necessity for determining r.]

CONTINUATION OF DR JONES'S PAPER

BY M. G. KENDALL

1. In his original version of the foregoing paper Dr Jones included a Part II on the estimation of mean-square error. His treatment was open to misunderstanding and his methods of estimating constants unnecessarily elaborate. The most unfortunate accident of his death prevented a discussion of these matters with him. The editors felt that his work was important enough to justify publication and the foregoing paper appears accordingly. In this continuation I have re-examined the work in his Second Part and have added some comments of my own. It was, the Editors felt, scarcely within their province to substitute this work for that of Dr Jones, and to publish it under his name, although some of the methods are due to him.

2. In the first place I wish to comment on Dr Jones's theorem 2 and its consequences. He has shown in his theorem 1 that the optimum distribution of sampling points is obtained when they are equidistant, and essentially his second theorem is concerned with the relative length of the two end-segments of the range of t as divided by the sampling points.

Write
$$\alpha = \frac{t_1 - T_0}{t_2 - t_1}. \tag{1}$$

Then since $p = e^{-q(t_2-t_1)}$ and $p_0 = e^{-q(t_1-T_0)}$, $\alpha = \log p_0/\log p$. From Dr Jones's equation (4·7) we have $p_0^2 p^{n-1} = e^{-nr}$ and thus it follows that

$$\alpha = \tfrac{1}{2} - \frac{n}{2}\left(1 + \frac{r}{\log p}\right). \tag{2}$$

Theorem 2 shows that
$$\tfrac{1}{2} \leqslant \alpha \leqslant 1. \tag{3}$$

Thus, as n becomes larger $1 + r/\log p$ must tend to zero, or p tends to e^{-r}. Furthermore, since $r = (T - T_0)q/n$, r must tend to zero (and p accordingly to unity) as n tends to infinity, *provided that* $(T - T_0)q$ *remains constant*. Since q is an (unknown) positive constant of the system, this is equivalent to the proviso that $T - T_0$ must remain constant.

3. I think Dr Jones's Table 1 tends to obscure this point. If we keep the interval $T - T_0$ constant, then as n increases p tends to unity, as we should expect from considerations of continuity. *Per contra*, for large n the interval increases as p becomes smaller. The increase in n for fixed p then does not correspond to a denser distribution of sample points but to the extension of the sampled strip. In such a case it seems an unnecessary refinement to discuss at great length the end intervals when the total interval is tending to infinity. Dr Jones may have been thinking here, not of his example of a strip of soil, but of meteorological variation which goes on indefinitely.

4. Let $(T - T_0)q = d$. Since p tends to unity with increasing n for fixed d we may write

$$1 - p = \frac{a_1}{n} + \frac{a_2}{n^2} + \frac{a_3}{n^3} + \frac{a_4}{n^4} + \ldots. \tag{4}$$

On substituting in
$$e^{-nr}(1-p)^2 = r^2 p^{n+1} \tag{5}$$

and identifying coefficients we find

$$a_1 = d, \quad a_2 = -\tfrac{1}{2}d^2, \quad a_3 = \frac{(2d+3)}{d+2}\frac{d^2}{12}, \quad a_4 = -\frac{d^5}{24(d+2)}. \tag{6}$$

Hence we find
$$1+\frac{r}{\log p}=\frac{a_2}{n^2}+\frac{1}{3}\frac{a_1^3}{dn^3}+O(n^{-3})$$
$$=-\frac{1}{6}\frac{d^2}{n^2}+O(n^{-3}). \tag{7}$$

Thus to order n^{-1}
$$\alpha=\frac{1}{2}+\frac{1}{12}\frac{d^2}{n}, \tag{8}$$

and so, as n tends to infinity α tends to $\frac{1}{2}$ in all cases where the interval d is fixed. It appears to me, therefore, that the general rule for large samples is to divide the interval into n equal parts and place one sample member at the middle of each. The discussion for small p is only appropriate when the sample is large and the interval is also large, in which case the location of the first sample member can scarcely be of much practical interest.

5. If we keep p fixed and let n tend to infinity, suppose that
$$1+\frac{r}{\log p}=\frac{b_1}{n}+\frac{b_2}{n^2}+\ldots. \tag{9}$$

Such an expansion is possible because, from (2), $1+r/\log p$ must tend to zero with increasing n in all cases. On substitution in (5) we find to order n^{-1}
$$\frac{p^{1+b_1}(\log p)^2}{(1-p)^2}\left(1-\frac{2b_1}{n}\right)=1-\frac{b_2}{n}\log p,$$

whence
$$b_1=-1+\frac{2}{\log p}\{\log(1-p)-\log|\log p|\}, \tag{10}$$
$$b_2=2b_1/\log p. \tag{11}$$

Thus
$$\alpha=1-\frac{1}{\log p}\{\log(1-p)-\log|\log p|\}. \tag{12}$$

It happens to be true that as p tends to unity this value of α tends to $\frac{1}{2}$.

6. It may be noted in passing that expansion by these methods (d fixed, n large) leads to the series
$$\alpha=\frac{1}{2}+\frac{r}{24}\frac{d}{d+2}+O(r^3), \tag{13}$$

which is not the same as Dr Jones's equation (4·10) unless we let d become large.

7. If the trend is constant or linear the mean-square error of our estimate for $\bar{Z}$ is the same as that for $\bar{X}$. We then require
$$E\left(\frac{1}{n}\Sigma x-\bar{X}\right)^2=\frac{1}{n^2}E(\Sigma x)^2-\frac{2}{n}E(\Sigma x\bar{X})+E(\bar{X})^2. \tag{14}$$

Now
$$E(\bar{X}^2)=\frac{1}{(T-T_0')^2}\int_{T_0}^{T}\int_{T_0}^{T}E\{X(t)\,X(u)\}\,dt\,du$$
$$=\frac{1}{(T-T_0')^2}\int_{T_0}^{T}\int_{T_0}^{T}\lambda e^{-q|t-u|}\,dt\,du$$
$$=\frac{2\lambda}{(T-T_0')q}-\frac{2\lambda(1-e^{-q(T-T_0')})}{(T-T_0')^2q^2}. \tag{15}$$

Using Dr Jones's (3·5) we have

$$\frac{2}{n}E(\Sigma x\bar{X}) = \frac{2}{n}\sum_j \frac{\lambda}{q}(2-p_0p^{j-1}-p_0p^{n-j})$$

$$= \frac{2\lambda}{nq}\left(2n-\frac{2p_0(1-p^n)}{1-p}\right)$$

$$= \frac{4\lambda}{(T-T_0)q}\left\{1-\frac{p(1-p^n)q(T-T_0)}{n^2(1-p)^2}\right\}, \tag{16}$$

$$E\left(\frac{1}{n}\Sigma x\right)^2 = \frac{1}{n^2}(\Sigma x^2+\Sigma x_ix_j) \quad (i\neq j)$$

$$= \frac{\mu}{n}+\lambda\{1+p+p^2+\ldots+p^{n-1}+p+1+p+\ldots+p^{n-2}+\ldots+p^{n-1}+p^{n-2}+\ldots+1\}$$

$$= \frac{\mu}{n}+\frac{\lambda(1+p)}{n(1-p)}-\frac{2\lambda p(1-p^n)}{n^2(1-p)^2}. \tag{17}$$

Hence, on substitution in (9), we find

$$S = E\left(\frac{1}{n}\Sigma x-\bar{X}\right)^2 = \frac{\mu}{n}+\lambda\left\{\frac{1+p}{n(1-p)}-\frac{2}{q(T-T_0)}\right\}$$

$$+2\lambda\left\{\frac{p(1-p^n)}{n^2(1-p)^2}-\frac{1-e^{-q(T-T_0)}}{(T-T_0)^2q^2}\right\}. \tag{18}$$

Since

$$\frac{p^{n+1}}{n^2(1-p)^2} = \frac{e^{-nr}}{n^2r^2} = \frac{e^{-q(T-T_0)}}{(T-T_0)q^2},$$

this simplifies slightly to

$$S = \frac{\mu}{n}+\lambda\left\{\frac{1+p}{n(1-p)}-\frac{2}{q(T-T_0)}\right\}+\frac{2\lambda}{n^2}\left\{\frac{p}{(1-p)^2}-\frac{n^2}{q^2(T-T_0)^2}\right\}. \tag{19}$$

8. Dr Jones, having reached this expression by a slightly different route, proceeded to argue that the last term was of order n^{-2} and could be neglected, whereas the second term was of order n^{-1}. Noting that $q(T-T_0)$ was of order $n\log 1/p$ he was led to the expression, for large n,

$$S = \frac{1}{n}\{\mu+\lambda F(p)\}, \tag{20}$$

where

$$F(p) = \frac{1+p}{1-p}+\frac{2}{\log p}. \tag{21}$$

He gave the attached Table 2 for $F(p)$ and emphasized that even for low values of p such as 0·01 the mean-square error (apart from the term in μ) was only half of the value for $p = 0$. He inferred that even a trace of correlation between successive sample members would seriously affect the mean-square error.

9. Now if we hold p constant and let n tend to infinity, formulae (20) and (21) result. But in doing so, as I have already pointed out, we are extending the interval to infinity, not concentrating the sample more densely. In these circumstances it is not so surprising that small

correlations should affect the mean-square error substantially. It appears to me to be more relevant to the problem to consider the limiting form of (19) when n tends to infinity but $d = q(T - T_0)$ remains fixed. In such a case we get quite different results.

10. Consider the first term in braces on the right-hand side of (19).

Substituting from (4) we find

$$\frac{1+p}{n(1-p)} - \frac{2}{d} = \left(2 - \frac{a_1}{n} - \frac{a_2}{n^2} - \frac{a_3}{n^3} - \ldots\right) \Big/ \left(a_1 + \frac{a_2}{n} + \frac{a_3}{n^2} + \ldots\right) - \frac{2}{d},$$

which, to order n^{-2}, reduces to

$$\frac{(d+3)\,d}{6(d+2)\,n^2}, \tag{22}$$

which is of order n^{-2}, not n^{-1}.

Similarly for the last expression in (19)

$$\frac{p}{n^2(1-p)^2} - \frac{1}{d^2} = -\frac{d}{12(d+2)\,n^2}. \tag{23}$$

To order n^{-1} we then have merely

$$S = \frac{\mu}{n}, \tag{24}$$

or, to order n^{-2},

$$S = \frac{\mu}{n} + \frac{\lambda d}{6n^2}. \tag{25}$$

Table 2 (*see text*)

p	$F(p)$	p	$F(p)$	p	$F(p)$
0·0	1·000	0·06	0·417	0·35	0·172
$0{\cdot}0^51$	0·855	0·07	0·398	0·40	0·151
0·001	0·712	0·08	0·382	0·45	0·137
0·005	0·633	0·09	0·367	0·50	0·115
0·01	0·586	0·10	0·354	0·60	0·085
0·02	0·530	0·15	0·299	0·70	0·059
0·03	0·491	0·20	0·257	0·80	0·037
0·04	0·462	0·25	0·224	0·90	0·018
0·05	0·438	0·30	0·196	1·00	0·000

11. The appearance of the term in μ is to be expected; it represents the variance of superposed error of observation. But the result that the remaining part of the mean-square error is of order n^{-2} is at first sight surprising. One is so accustomed in statistical work to a sampling variance of order n^{-1} that anything of lower order requires some explanation.

It is here that we must remember that our problem is not the determination of the mean-square error or the sampling variance of n independent observations. On the contrary, under our assumptions, the correlation between neighbouring sample members tends to unity with increasing n. The variation of such members among themselves is of order n^{-2} as may be verified from (17), neglecting the term in μ. Since we choose our function $\bar{X}$ so as to fit these observations as closely as possible it is not, after all, surprising that the average difference square of $\frac{1}{n}\Sigma(x)$ and $\bar{X}$ is also of order n^{-2}.

12. There is one essential discontinuity in the situation, however, which is worth noticing. If $q(T-T_0)$ remains finite then if n tends to infinity, p must tend to unity and the series (apart from error of observation) is continuous. However large q may be this is true. But if q is itself infinite the series is not necessarily continuous, the observations are independent and we revert to the usual situation of n independent observations. Thus for any q, however large, (25) is correct; but for q infinite it is not. Looking back to (19) we see that (since p in this case is zero)

$$S = \frac{\mu}{n} + \frac{\lambda}{n},$$

which is what we should expect.

13. From (19) we see that (without approximation) the mean-square error S depends on μ, λ and p; on the known quantities $T-T_0$ and n; and on q, which is known when p is known and the interval between successive observations is known. In practice we may require to estimate μ, λ and p. In the limiting case which Dr Jones considered, leading to (20) we also require μ, λ and p. In the limiting case which I am considering, leading to (25), we require only μ and λ.

Dr Jones proceeded by dismissing μ and considering the estimation of λ and p from the expectations of powers of differences $\Sigma(x_j - x_{j-1})^2$. This is really equivalent to using the serial correlations of the observations.

We have, since $E(x) = 0$,

$$E(x^2) = \mu + \lambda, \quad E(x_j x_{j+1}) = \lambda p, \quad E(x_j x_{j+2}) = \lambda p^2, \text{ etc.}$$

If we take the observed serial covariances of the observations (freed from trend) as estimators of the corresponding expectations we then have, if the serial correlations are r_1, r_2, etc.

$$p = \frac{r_2}{r_1}, \tag{26}$$

$$\lambda = \frac{r_1^2}{r_2} \operatorname{var} x, \tag{27}$$

$$\mu = \operatorname{var} x \left(1 - \frac{r_1^2}{r_2}\right). \tag{28}$$

These seem to me to be the simplest equations of estimation which one is likely to find. If we may assume that $\mu = 0$ we have the more reliable forms

$$p = r_1, \tag{29}$$

$$\lambda = \operatorname{var} x. \tag{30}$$

These equations are the usual ones for estimating constants in a Markoff series.

14. One final comment. The relative simplicity of Dr Jones's result that the optimum distribution of sample points is equidistantly along the interval is a little deceptive. It is natural to consider the more general problem when the autocorrelations along the series are not necessarily decaying according to an exponential law, but have any form permissible for a continuous random process. An intuitive approach might suggest that the weights and distances of the observations should be equal because there is no obvious reason to the contrary; but any conclusion of this kind would be quite wrong. If we denote the autocorrelation

function of the series by $\rho(t)$ and the sample points are $t_1 \dots t_n$ in that order the general equations corresponding to Dr Jones's (3·11) and (3·12) are

$$\left.\begin{aligned}
\left(1+\frac{\mu}{\lambda}\right)v_1 + v_2\rho(t_2-t_1) + \dots + v_n\rho(t_n-t_1) &= E\{x(t_1)\,\bar{X}\},\\
v_1\rho(t_2-t_1) + v_2\left(1+\frac{\mu}{\lambda}\right) + \dots + v_n\rho(t_n-t_2) &= E\{x(t_2)\,\bar{X}\},\\
\dots\dots\dots

THE PRECISION OF OBSERVED VALUES OF SMALL FREQUENCIES

By J. B. S. HALDANE, F.R.S.

In recent genetical work numerous observers have recorded the frequencies of rare events, notably mutations. It has been realized that it is misleading to state the observed frequencies with their standard errors, since the distribution is decidedly skew. Various devices have been suggested to avoid this difficulty. But so far as I know it has not been pointed out that, when the frequency is small, its cube root is almost normally distributed. This will be proved and applied to actual observations.

Let a rare event be observed in a out of n trials, where n is much greater than a^2. Let x be the true value of the frequency, whose observed value is $p = a/n$. Let the *a priori* distribution of x be

$$dF = \phi(x)\,dx.$$

Let the probability distribution, after the observation has been made, be

$$dF = f(x)\,dx,$$

and let $x = y^3$.

Then for given values of n and x, the probability of a is

$$\binom{n}{a} x^a (1-x)^{n-a}.$$

Hence for given values of n and a, the distribution of x is

$$dF = f(x)\,dx = \frac{x^a(1-x)^{n-a}\phi(x)\,dx}{\int_0^1 x^a(1-x)^{n-a}\phi(x)\,dx}.$$

If we assume that all values of x are equiprobable, $\phi(x) = 1$, and

$$\bar{x} = \frac{(n+1)!}{a!\,(n-a)!}\int_0^1 x^{a+1}(1-x)^{n-a}\,dx = \frac{a+1}{n+2}.$$

This value should of course be a/n. As I have previously remarked (Haldane, 1932) and as Jeffreys (1948) has shown in greater detail, the assumption that $\phi(x) = 1$ introduces a bias. It is also contrary to common sense. If we are trying to estimate a mutation rate, we know *a priori* that it will almost certainly be less than 10^{-3} and greater than 10^{-20}. In a particular case we might perhaps guess that such a rate would be about as likely to lie between 10^{-5} and 10^{-6} as between 10^{-6} and 10^{-7}. In other words, when x is small it is more nearly true that all values of $\log x$ are equiprobable than that all values of x are equiprobable. This would imply that $\phi(x) = c/x$ in the region considered. However, this cannot continue to be true when x is sufficiently small. If we wished to state a plausible general form for the *a priori* distribution of x it might be somewhat as follows:

$$F = k \quad (x = 0),$$

$$dF = \frac{C}{(x+\epsilon)(1+\epsilon-x)}\,. \quad (0 < x < 1),$$

$$F = k \quad (x = 1),$$

where k is some number less than $\frac{1}{2}$ expressing the possibility that x may prove to be zero or unity,

$$C = \frac{(1-2k)(1+2\epsilon)}{2\log(1+\epsilon^{-1})},$$

and ϵ is a very small number, perhaps of the order of 10^{-100}, expressing the fact that exceedingly rare events are relatively infrequent. If the universe is finite in space and in time, and if there is a minimum time in which an event can occur, it might imply that there is no sense in discussing events which have no appreciable probability of ever occurring.

For practical purposes, however, so long as we know that a exceeds zero, and is less than n, that is to say, that the event considered is possible and so is its converse, we can take $\phi(x) = \dfrac{C}{x(1-x)}$ without appreciable error. We then have

$$dF = \frac{(n-1)!}{(a-1)!\,(n-a-1)!}x^{a-1}(1-x)^{n-a-1}\,dx. \tag{1}$$

This is a Pearsonian Type I distribution, and

$$\overline{x^r} = \frac{(n-1)!\,(a+r-1)!}{(n+r-1)!\,(a-1)!}.$$

Thus $\bar{x} = a/n$, as it should be, $\overline{x^2} = \dfrac{a(a+1)}{n(n+1)}$, etc. When an^{-1} is small, this approximates very closely to the Type III distribution

$$dF = \frac{e^{-nx}x^{a-1}}{(a-1)!}dx. \tag{2}$$

Now Wilson & Hilferty (1931) showed that the cube root of χ^2 is almost normally distributed; and the same transformation will almost normalize many Type III distributions.

The standard form of this type, referred to its mode, is

$$dF = C\left(1+\frac{x}{a}\right)^{\gamma a}\epsilon^{-\gamma x}\,dx.$$

It is more convenient to change the origin to the point where the probability becomes zero, and write

$$dF = \frac{\gamma^c x^{c-1}\,dx}{\Gamma(c)\,e^{\gamma x}},$$

where $$c = 1+p = 1+\gamma a = 4/\beta_1.$$

$\kappa_r = (r-1)!\,c\gamma^{-r}$, so the mean is $c\gamma^{-1}$ and the moments about it are

$\mu_2 = c\gamma^{-2}$, $\mu_4 = (3c^2+6c)\gamma^{-4}$, $\mu_6 = (15c^3+130c^2+120c)\gamma^{-6}$, $\mu_8 = (105c^4+\ldots)\gamma^{-8}$,
$\mu_3 = 2c\gamma^{-3}$, $\mu_5 = (20c^2+24c)\gamma^{-5}$, $\mu_7 = (210c^3+924c^2+720c)\gamma^{-7}$, $\mu_9 = (2520c^4+\ldots)\gamma^{-9}$.

Let $x = c\gamma^{-1}+z$, so that $\overline{z^r} = \mu_r$ and $y = (\gamma x/c)^{\frac{1}{3}}$. Then

$$y = \left(1+\frac{\gamma z}{c}\right)^{\frac{1}{3}}$$

and
$$\overline{y^r} = 1+\tfrac{1}{3}r\frac{\gamma\bar{z}}{c}+\binom{\frac{1}{3}r}{2}\frac{\gamma^2\bar{z}^2}{c^2}+\binom{\frac{1}{3}r}{3}\frac{\gamma^3\bar{z}^3}{c^3}+\ldots$$

$$= 1+\frac{r(r-3)}{18c}+\frac{r(r-1)(r-3)(r-6)}{2(18c)^2}+\frac{r^2(r-3)^2(r-6)(r-9)}{6(18c)^3}$$

$$+\frac{r(r-3)(r-6)(r-9)(r-12)(5r^3-30r^2+15r+18)}{120(18c)^4}+O(c^{-5}).$$

Or, putting $t = \frac{1}{9c}$,

$$\begin{aligned}
\overline{y} &= 1 - t + \tfrac{10}{3}t^3 + \tfrac{11}{3}t^4 + O(t^5),\\
\overline{y^2} &= 1 - t + t^2 + \tfrac{7}{3}t^3 - \tfrac{28}{3}t^4 + O(t^5),\\
\overline{y^3} &= 1,\\
\overline{y^4} &= 1 + 2t - 3t^2 + \tfrac{10}{3}t^3 + \tfrac{41}{3}t^4 + O(t^5),\\
\overline{y^5} &= 1 + 5t - 5t^2 + \tfrac{25}{3}t^3 + \tfrac{14}{3}t^4 + O(t^5),\\
\overline{y^6} &= 1 + 9t.
\end{aligned}$$

Hence the cumulants of the distribution of y are

$$\left.\begin{aligned}
\kappa_1 &= 1 - t + \tfrac{10}{3}t^3 + \tfrac{11}{3}t^4 + O(t^5),\\
\kappa_2 &= t - \tfrac{13}{3}t^3 - 10t^4 + O(t^5),\\
\kappa_3 &= 4t^3 + 16t^4 + O(t^5),\\
\kappa_4 &= -2t^3 - 16t^4 + O(t^5),\\
\kappa_5 &= 8t^4 + O(t^5),\\
\kappa_6 &= -55t^4 + O(t^5),
\end{aligned}\right\} \qquad (3)$$

or

$$\begin{aligned}
\sigma &= \frac{1}{(9c)^{\frac{1}{2}}}\left[1 - \frac{13}{6(9c)^2} - \frac{5}{(9c)^3} + O(c^{-4})\right],\\
\gamma_1 &= \frac{4}{(9c)^{\frac{1}{2}}}\left[1 + \frac{4}{9c} + O(c^{-2})\right],\\
\gamma_2 &= \frac{-2}{9c}\left[1 + \frac{8}{9c} + O(c^{-2})\right],\\
\gamma_3 &= \frac{8}{(9c)^{\frac{1}{2}}} + O(c^{-\frac{1}{2}}),\\
\gamma_4 &= \frac{-55}{9c} + O(c^{-2}).
\end{aligned}$$

Thus provided $9c$, or $9(1+p)$, is large, the approximation to normality, up to the sixth moment, is satisfactory. But it is of no value when p is negative, that is to say, the curve is J-shaped. (Here p is of course the parameter used in specifying Type III distributions, and not the observed frequency value.)

To apply these formulae to the distribution of y, we have only to put $a = c$, and to multiply κ_r by $p^{\frac{1}{2}r}$. We thus find

$$\left.\begin{aligned}
\kappa_1 &= p^{\frac{1}{2}}[1 - \tfrac{1}{9}a^{-1} + \tfrac{10}{2187}a^{-3} + O(a^{-4})], \quad \kappa_3 = p[\tfrac{4}{729}a^{-3} + O(a^{-4})],\\
\kappa_2 &= p^{\frac{1}{2}}[\tfrac{1}{9}a^{-1} - \tfrac{13}{2187}a^{-3} + O(a^{-4})], \quad \kappa_4 = p^{\frac{1}{2}}[\tfrac{2}{729}a^{-3} + O(a^{-4})], \quad \text{etc.}
\end{aligned}\right\} \qquad (4)$$

Thus $\sigma = p^{\frac{1}{2}}/3a^{\frac{1}{2}}$, $\gamma_1 = \frac{4}{27}a^{-\frac{1}{2}}$, $\gamma_2 = \frac{2}{9}a^{-1}$, all approximately.

The terms involving a^{-3} in the mean and standard error may be safely neglected in practice. Even when $a = 1$, the former is only 0·013 of the standard error. If we take (1) as our distribution of x, a term of order n^{-1} must be added to those of order a^{-3}. This also can be safely neglected.

Thus we find that y is almost normally distributed with mean $(1 - \frac{1}{9}a^{-1})\,p^{\frac{1}{2}}$, and standard deviation $p^{\frac{1}{2}}/3a^{\frac{1}{2}}$. For example, if $n = 1000$, $a = 8$, $p = 0{\cdot}008$, x is by no means normally distributed about 0·008, for $\beta_1 = 0{\cdot}5$ and $\beta_2 = 3{\cdot}75$. But y is very nearly normally distributed

about $0{\cdot}2 \times \frac{71}{72}$ or 0·1972 with standard deviation $\frac{1}{120}$ or 0·0083, with $\beta_1 = 0{\cdot}00004$, and $\beta_2 = 3{\cdot}028$. The method of Haldane (1938) would give an even better fit if $a > 10$.

Two examples will be given showing how the method can be actually used.

Muller (1928, p. 311) found 13 lethal genes in 1034 X-chromosomes of flies kept at 27° C., and 5 in 840 X-chromosomes of flies kept at 19·5° C. Thus corrected values of y_1 and y_2 are:—

$$y_1' = \left(\frac{13}{1034}\right)^{\frac{1}{3}}\left(\frac{1}{1-9\times 13}\right) = 0{\cdot}23054 \pm 0{\cdot}02150, \quad y_2' = 0{\cdot}17720 \pm 0{\cdot}02702.$$

$y_1' - y_2' = 0{\cdot}05334$, which is 1·55 times its standard error of 0·03452. The difference is therefore rather more significant than Muller, who used the usual formula, believed.

Again Muller (1940) obtained 7 translocations in 3366 flies with a dose of 375 r., and 56 in 2223 flies with a dose of 1500 r. The question at issue was as follows: 'the frequency may be proportional to the dosage, to its $\frac{3}{2}$th power or to its square. With which, if any, of these hypotheses are the observed results consistent?'

$$y_1' = 0{\cdot}125616 \pm 0{\cdot}016081, \quad y_2' = 0{\cdot}29256 \pm 0{\cdot}01306.$$

We therefore compare y_2' with

$$2^{\frac{1}{3}}y_1' = 0{\cdot}19940 \pm 0{\cdot}02559, \quad 2y_1' = 0{\cdot}25123 \pm 0{\cdot}03216, \quad 2^{\frac{4}{3}}y_1' = 0{\cdot}31653 \pm 0{\cdot}04052.$$

The differences are respectively 3·25, 1·19 and 0·56 times their standard errors, so either of the latter two hypotheses is admissible.

It is perhaps worth remarking that, if the emendation of the classical inverse probability distribution be rejected, and the calculation made according to Bayes's hypothesis, the cube root of the frequency is still almost normally distributed. It is also true that if the frequency of a rare event is estimated by the method described by Haldane (1945) when the observations cease when a fixed number m of rare events have occurred, the estimated frequency being $(m-1)/(n-1)$, where n is the total number of observations, the cube root of the estimate is almost normally distributed. Here too the cube root may be used with advantage in comparing different estimates.

I have to thank Prof. E. S. Pearson for valuable criticism.

Summary

When an event is rare, the distribution of the cube root of the frequency round the cube root of the estimate is much more nearly normal than the distribution of the true frequency round the estimate.

REFERENCES

Haldane, J. B. S. (1932). A note on inverse probability. *Proc. Camb. Phil. Soc.* **28**, 55–61.

Haldane, J. B. S. (1938). The approximate normalization of a class of frequency distributions. *Biometrika*, **29**, 392–404.

Haldane, J. B. S. (1945). On a method of estimating frequencies. *Biometrika*, **33**, 222–5.

Jeffreys, H. (1948). *Theory of Probability*. Oxford University Press.

Muller, H. J. (1928). The measurement of gene mutation rate in *Drosophila*, its high variability and its dependence on temperature. *Genetics*, **13**, 274–357.

Muller, H. J. (1940). An analysis of the process of structural change in chromosomes of *Drosophila*. *J. Genet.* **40**, 1–66.

Wilson, E. B. & Hilferty, M. M. (1931). The distribution of Chi-square. *Proc. Nat. Acad. Sci., Wash.*, **17**, 684–88.

NOTE ON PROFESSOR HALDANE'S PAPER REGARDING THE TREATMENT OF RARE EVENTS

By E. S. PEARSON

In the preceding paper Prof. Haldane has suggested a method of handling certain problems involving the occurrence of rare events, by introducing the cube-root transformation. His method of attack involves the use of the concept of inverse probability, so that his final, closely normal distribution is the posterior probability distribution of y, the cube root of the unknown probability x of the occurrence of the event. From this point of view x, and therefore y, are continuous variables. If we attack the problem without an appeal to inverse probability, we are concerned with the probability distribution of a, the observed sample frequency, for a given x. Since a is a discontinuous random variable which, in the problems considered, may only assume the first few integer values $0, 1, 2, \ldots$, the cube-root transformation would here clearly introduce some awkward problems of discontinuity. It seems of interest to consider how the examples which Haldane gives could be dealt with by the direct method without the need of substituting approximate standard errors or, indeed, of assuming that a very skew distribution is normal.

In both Haldane's examples we are concerned with a comparison of the results of two experiments. In a first sample of size n_1 an event occurs on a_1 occasions, the chance of occurrence being x_1; similarly for the second sample we have a_2, n_2 and x_2. We then ask whether the results are consistent with the hypothesis that $x_1 = \kappa x_2$, where, for the first example $\kappa = 1$ and for the second has to be taken successively as $\frac{1}{4}$, $(\frac{1}{4})^{\frac{1}{2}}$ and $(\frac{1}{4})^2$.

The assumption which I shall make is that in the problems considered x and $1/n$ are sufficiently small to justify the use of the Poisson series in place of the binomial expansion $(\overline{1-x}+x)^n$. If this is the case and we write

$$m_i = n_i x_i \quad (i = 1, 2),$$

the chance of the observed event may be written

$$p(a_1, a_2 \mid m_1, m_2) = e^{-(m_1+m_2)}\frac{m_1^{a_1} m_2^{a_2}}{a_1! a_2!} = \frac{e^{-\mu}\mu^{a_1+a_2}}{(a_1+a_2)!} \times \frac{(a_1+a_2)!}{a_1! a_2!}\lambda^{a_1}(1-\lambda)^{a_2}, \tag{1}$$

where

$$\mu = m_1 + m_2, \quad \lambda = \frac{m_1}{m_1+m_2}. \tag{2}$$

For a fixed value of $r = a_1 + a_2$, the relative distribution of a_1 follows a binomial, $(\overline{1-\lambda}+\lambda)^r$, where λ is specified by hypothesis. Thus if the hypothesis is that $m_1 = m_2$, then $\lambda = \frac{1}{2}$, but in general

$$\lambda = \frac{n_1 x_1}{n_1 x_1 + n_2 x_2} = \frac{n_1 \kappa}{n_1 \kappa + n_2}, \tag{3}$$

where κ is the ratio of the true or hypothetical chances, x_1/x_2.

As in the case of the more general problem of the analysis of 2×2 tables, where the conditional distribution is a hypergeometric series, there may be some difference of opinion on the way in which the result is used.

(1) We may consider that the whole answer lies in the conditional distribution of a_1 (or a_2) for fixed r, in the sense that we need only ask whether the partition of the observed events

into a_1 occurring in the first sample and a_2 in the second is consistent with our hypothesis as to κ. The answer is obtained in terms of probability by summing the tail terms of the binomial $(\overline{1-\lambda}+\lambda)^r$. This approach appears a natural one to take when two treatments have been randomly assigned among n_1+n_2 individuals.

(2) We may wish to set the observed result against the two-dimensioned distribution of a_1 and a_2 obtainable in unrestricted sampling, i.e. without the condition that $r = a_1+a_2$ is fixed. This distribution depends on the value of $\mu = m_1+m_2$ which is not specified by the hypothesis, but it is possible to obtain upper limits to significance levels; this problem was considered by Przyborowski & Wilenski (1939) for the case $\lambda = \frac{1}{2}$, and further tables for some other special cases were circulated during the war within the Ministry of Supply (Barnard, 1944; Allinson, 1944).

(3) Still avoiding the condition that r is fixed, we may note that for unrestricted sampling

$$u = \frac{a_1 - r\lambda}{\sqrt{\{r\lambda(1-\lambda)\}}} \tag{4}$$

is a random variable which, on the hypothesis tested, has a zero expectation and unit variance. Its sampling distribution is composite, being the sum of a number of binomials combined with weights depending on the unknown value of $\mu = m_1+m_2$. However, if λ is not too different from $\frac{1}{2}$, this distribution will not be very far from the normal even when dealing with small frequencies. A calculation of the ratio u will therefore often provide the broad answer needed in practice.

These points are illustrated on Haldane's examples.

Example 1. $a_1 = 13$, $n_1 = 1034$; $a_2 = 5$, $n_2 = 840$; $r = 18$. In this problem, the hypothesis tested is that $x_1 = x_2$, so that $\kappa = 1$ and $\lambda = n_1/(n_1+n_2) = 0{\cdot}5518$. On the assumption that the flies were randomly divided into the two temperature groups, the null hypothesis is that 18 out of the 1874 would have produced progeny of a type which showed they carried a lethal gene at whichever temperature they were kept. Using the binomial approximation to the hypergeometric, a_1 would assume values of $0, 1, 2, \ldots, 18$ with probabilities given by the expansion of $(0{\cdot}4482+0{\cdot}5518)^{18}$. The chance that $a_1 \geqslant 13$ is given by the Incomplete Beta Function Ratio

$$I_\lambda(a_1, r-a_1+1) = I_{0{\cdot}5518}(13, 6) = 0{\cdot}1107.$$

An approximation to this chance can be obtained from the integral under the normal curve having

$$\text{Mean} = r\lambda = 9{\cdot}932, \quad \text{S.D.} = \sqrt{\{r\lambda(1-\lambda)\}} = 2{\cdot}110,$$

using the correction for continuity. We then get the ratio $(12{\cdot}5-9{\cdot}932)/2{\cdot}110 = 1{\cdot}217$, corresponding to a chance of 0·1118.

If we take the approach of (3) above, avoiding restriction to the conditional set $r = 18$, then a correction for continuity is not appropriate and we find

$$u = (13-9{\cdot}932)/2{\cdot}110 = 1{\cdot}45,$$

a ratio to be compared with Haldane's 1·55, obtained from the posterior distribution of $x_1^{\frac{1}{2}} - x_2^{\frac{1}{2}}$.

Example 2. Here $a_1 = 7$, $n_1 = 3366$; $a_2 = 56$, $n_2 = 2223$; $r = 63$. Three hypotheses are examined, (*a*), (*b*) and (*c*), and as two of these involve the assumption that the chance x of translocation varies with dose, the randomization approach to the conditional distribution is less clear.* I think I should here base my conclusions on the values of the ratio u. Relevant

* In this case, with $a_2 = 56$, we are beyond the range of the Tables of the Incomplete Beta Function.

figures are given below, the means and standard deviations being for the binomials $(\overline{1-\lambda}+\lambda)^{63}$. It will be seen that the values of $(a_1-r\lambda)/\sqrt{\{r\lambda(1-\lambda)\}}$ are not very different from Haldane's ratios, and that the same conclusions would be reached using either method of approach.

	Hypothesis	Mean $a_1\|r$	S.D. of $a_1\|r$	$u=(a_1$-mean)/S.D.	Haldane's ratio
(*a*)	$\kappa=\frac{1}{4}$, $\lambda=0{\cdot}2746$	17·300	3·542	−2·91	−3·24
(*b*)	$\kappa=\frac{1}{8}$, $\lambda=0{\cdot}1591$	10·026	2·904	−1·04	−1·19
(*c*)	$\kappa=\frac{1}{16}$, $\lambda=0{\cdot}0865$	5·447	2·231	0·70	0·56

To sum up:

(1) if conditions justify us in using the Poisson series to represent the binomial distribution of a, given x, which will generally be the case with 'rare' events,

(2) if we are content to base our answer on the conditional distribution of a_1 for fixed $r=a_1+a_2$,

the direct method of attack in this two-sample problem provides a test involving a binomial distribution. The exact answer in terms of a significance level requires the calculation of the sum of tail terms of the binomial, which can be obtained directly from the Tables of the Incomplete Beta Function if both a_1 and $a_2+1\leqslant 50$. Alternatively, if we do not wish to restrict variation to the conditional set, we may obtain a less precise answer by referring $u=(a_1-r\lambda)/\sqrt{\{r\lambda(1-\lambda)\}}$ to the normal integral, a procedure which will be satisfactory if λ is not too far from 0·5.

Haldane's solution with its continuous normal distribution is obviously attractive, but it has involved the introduction of the theory of inverse probability.

REFERENCES

ALLINSON, V. A. (1944). Technical Report Q.C./R/21 (Ministry of Supply).
BARNARD, G. A. (1944). Technical Report Q.C./R/18 (Ministry of Supply).
HALDANE, J. B. S. (1948). *Biometrika*, **35**, 297.
PRZYBOROWSKI, J. & WILENSKI, H. (1939). *Biometrika*, **31**, 313.

A FURTHER NOTE ON THE MEAN DEVIATION

By H. J. GODWIN, *University College of Swansea*

1. *Introduction.* In a previous paper (1945), I obtained the distribution of the estimate of mean deviation obtained from samples from a normal population; the method of derivation was an algebraic transformation of the sample space. In the present note I explain the geometrical significance of the result, and obtain a method of computing the moments of the distribution. Finally, I discuss various approximations to the distribution.

2. *Geometrical prologue.* The fundamental geometrical entity which will occur here is the regular simplex in k dimensions—a figure whose vertices are $(k+1)$ points each equidistant from the remainder. (Particular cases are the equilateral triangle and the regular tetrahedron.) These points lie on a hypersphere whose centre is the centroid of the simplex. Any k of the points form a regular simplex of $(k-1)$ dimensions; the join of the centroid of this to the centroid of the whole passes through the remaining vertex and is perpendicular to the space of the $(k-1)$-dimensional simplex. Since the centroid divides the perpendicular from a vertex to the opposite face in the ratio $k:1$, the angle made with each other by the (equally inclined) lines from the centroid to the vertices is $\arccos(-1/k)$. If the length of side of the simplex is a, then each vertex is at a distance $a\left(\frac{k}{2(k+1)}\right)^{\frac{1}{2}}$ from the centroid and, by a suitable choice of axes through the centroid, the vertices are

$$-\frac{a}{\sqrt{\{2k(k+1)\}}}, \quad -\frac{a}{\sqrt{\{2(k-1)k\}}}, \quad \ldots, \quad -\frac{a}{\sqrt{\{2(k-r+1)(k-r+2)\}}},$$
$$a\sqrt{\left(\frac{k-r}{2(k-r+1)}\right)}, \quad 0, 0, \ldots, 0 \quad (r = 0, 1, \ldots, k). \quad (1)$$

The vertices lie in sets of k on the $(k+1)$ bounding hyperplanes

$$-\sqrt{\left(\frac{2}{k(k+1)}\right)}x_1 - \sqrt{\left(\frac{2}{k(k-1)}\right)}x_2 - \ldots - \sqrt{\left(\frac{2}{(k+1-r)(k+2-r)}\right)}x_r$$
$$+\sqrt{\left(\frac{2(k-r)}{k+1-r}\right)}x_{r+1} + \frac{a}{k+1} = 0 \quad (r = 0, 1, \ldots, k). \quad (2)$$

The defining relations for the interior of the simplex are that the left-hand sides of (2) should be not less than zero.

We now find the value of the integral of $\exp\{-\frac{1}{2}(x_1^2+x_2^2+\ldots+x_k^2)\}$ taken through the interior of the simplex. Let this be $E(k, a)$. By joining the centroid of the simplex to the edges of any one of the $(k-1)$-dimensional simplexes in the bounding hyperplanes we obtain $(k+1)$ equal regions and the integral of $\exp\{-\frac{1}{2}(x_1^2+x_2^2+\ldots+x_k^2)\}$ through such a region is $E(k, a)/(k+1)$. A section of a region by a hyperplane parallel to the base hyperplane and at a perpendicular distance x from the centroid is a simplex of side $x\sqrt{\{2k(k+1)\}}$. Hence

$$\frac{E(k,a)}{k+1} = \int_0^{a/\sqrt{\{2k(k+1)\}}} E[k-1, x\sqrt{\{2k(k+1)\}}]\, e^{-\frac{1}{2}x^2}\, dx.$$

It follows, by induction, that $E(k, a) = \sqrt{(k+1)}\, G_k(a/\sqrt{2})$, (3)

where the function G_k is as defined in my 1945 paper. [This is, perhaps, the appropriate place in which to remark that this function is related to a function defined by McKay (1935) in dealing with the distribution of deviations from the greatest observation in a sample; in fact,

his $F_n(x) = \sqrt{n}\, G_{n-1}(nx)/(2\pi)^{\frac{1}{2}(n-1)}$. I am indebted to Dr H. O. Hartley for bringing this to my notice.]

From (3) it follows that $G_k(\infty) = \dfrac{1}{\sqrt{(k+1)}} E(k,\infty) = \dfrac{(2\pi)^{\frac{1}{2}k}}{\sqrt{(k+1)}}$.

3. *Distribution of the mean deviation.* Let a sample of n from the normal population (supposed, without loss of generality, to have zero mean and unit variance) be $x_1, x_2, \ldots, x_n$. Let the mean of the sample be $\bar{x}$, and let $d_i = x_i - \bar{x}$ $(i = 1, 2, \ldots, n)$. Consider the case in which k of the d's are negative and the rest positive. Then $1 \leqslant k \leqslant n-1$. By suitably numbering the d's the negative ones may be taken as $d_1, \ldots, d_k$; the mean deviation of the sample is then

$$m = n^{-1}[-d_1 - d_2 - \ldots - d_k + d_{k+1} + \ldots + d_n].$$

We now transform the sample space of the x's by means of the matrix equation $Y = TX$, where X, Y are column vectors $(x_1, \ldots, x_n)$, $(y_1, \ldots, y_n)$ and T is the orthogonal matrix:

k columns

$$\left.\begin{array}{cccc}
\frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\
-\frac{n-k}{\sqrt{\{kn(n-k)\}}} & -\frac{n-k}{\sqrt{\{kn(n-k)\}}} & \cdots & -\frac{n-k}{\sqrt{\{kn(n-k)\}}} \\
-\frac{k-1}{\sqrt{\{k(k-1)\}}} & \frac{1}{\sqrt{\{k(k-1)\}}} & \cdots & \frac{1}{\sqrt{\{k(k-1)\}}} \\
0 & -\frac{k-2}{\sqrt{\{(k-1)(k-2)\}}} & \cdots & \frac{1}{\sqrt{\{(k-1)(k-2)\}}} \\
\cdots\cdots & \cdots\cdots & \cdots & \cdots\cdots \\
0 & 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\
0 & 0 & \cdots & 0 \\
0 & 0 & \cdots & 0 \\
\cdots\cdots & \cdots\cdots & \cdots & \cdots\cdots \\
0 & 0 & \cdots & 0
\end{array}\right\}\begin{array}{l} k-1 \text{ rows} \\ \\ \\ \\ \\ \\ n-k-1 \text{ rows} \\ \\ \\ \\ \end{array}$$

$n-k$ columns

$$\left.\begin{array}{ccc}
\frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\
\frac{k}{\sqrt{\{kn(n-k)\}}} & \frac{k}{\sqrt{\{kn(n-k)\}}} & \cdots & \frac{k}{\sqrt{\{kn(n-k)\}}} \\
0 & 0 & \cdots & 0 \\
0 & 0 & \cdots & 0 \\
\cdots\cdots & \cdots\cdots & \cdots & \cdots\cdots \\
0 & 0 & \cdots & 0 \\
\frac{n-k-1}{\sqrt{\{(n-k)(n-k-1)\}}} & -\frac{1}{\sqrt{\{(n-k)(n-k-1)\}}} & \cdots & -\frac{1}{\sqrt{\{(n-k)(n-k-1)\}}} \\
 & \frac{n-k-2}{\sqrt{\{(n-k-1)(n-k-2)\}}} & \cdots & -\frac{1}{\sqrt{\{(n-k-1)(n-k-2)\}}} \\
\cdots\cdots & \cdots\cdots & \cdots & \cdots\cdots \\
0 & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}}
\end{array}\right\}\begin{array}{l} k-1 \text{ rows} \\ \\ \\ \\ \\ n-k-1 \text{ rows} \\ \\ \\ \\ \end{array}$$

Then $y_1 = \sqrt{n}\,\bar{x}$, and $y_2 = \frac{1}{2}n^{\frac{1}{2}}m/\sqrt{\{k(n-k)\}}$. Since $X = T'Y$, T' being the transposed matrix of T, the d's can easily be expressed in terms of the y's, and this gives, using equations (2), that the appropriate region in the space of the co-ordinates $y_3, \ldots, y_{k+1}$ is a regular simplex of side $y_2\{2k(n-k)/n\}^{\frac{1}{2}} = \frac{1}{2}nm$, and in the space of the co-ordinates $y_{k+2}, \ldots, y_n$, a regular simplex also of side $\frac{1}{2}nm$. The frequency function of m is $f(m)\,dm$, which is the integral of $2\pi^{-\frac{1}{2}n}\exp\{-\frac{1}{2}(x_1^2+\ldots+x_n^2)\}$, taken through the region in which the mean deviation lies between m and $m+dm$. Since T is an orthogonal matrix, $(x_1^2+\ldots+x_n^2) = (y_1^2+\ldots+y_n^2)$, and we may integrate over the variable y_1 (which is independent of m) from $-\infty$ to ∞. The contribution to $f(m)\,dm$ from the region described above is

$$\frac{1}{(2\pi)^{\frac{1}{2}(n-1)}}e^{-\frac{1}{2}y_1^2}\,dy_2\sqrt{\{k(n-k)\}}\,G_{k-1}(\tfrac{1}{2}nm)\,G_{n-k-1}(\tfrac{1}{2}nm)$$
$$= \frac{n^{\frac{1}{2}}}{2(2\pi)^{\frac{1}{2}(n-1)}}\exp\left[-\frac{n^3m^2}{8k(n-k)}\right]G_{k-1}(\tfrac{1}{2}nm)\,G_{n-k-1}(\tfrac{1}{2}nm)\,dm.$$

Now, for a given value of k, there are ${}^nC_k$ ways of allocating the k negative d's and k may vary from 1 to $n-1$. Hence

$$f(m)\,dm = \frac{n^{\frac{1}{2}}}{2(2\pi)^{\frac{1}{2}(n-1)}}\sum_{k=1}^{n-1}{}^nC_k\exp-\left[\frac{n^3m^2}{8k(n-k)}\right]G_{k-1}(\tfrac{1}{2}nm)\,G_{n-k-1}(\tfrac{1}{2}nm)\,dm,$$

as obtained before (1945).

Exact descriptions of the sample space for small values of n may be of interest. For $n = 2$, m is constant on two straight lines both perpendicular to the lines $\bar{x}$ = constant. For $n = 3$, m is constant on a cylinder, whose axis is perpendicular to the planes $\bar{x}$ = constant, and whose cross-section is a regular hexagon (i.e. ${}^3C_1+{}^3C_2$ equal lines). For $n = 4$, m is constant, for any space $\bar{x}$ = constant, on the surface of a cube of side $4m$, whose corners have been 'filed off' to give equilateral triangles of side $2\sqrt{2}\,m$. The surface thus consists of 6 (i.e. ${}^4C_2$) squares and 8 (i.e. ${}^4C_1+{}^4C_3$) equilateral triangles.

4. *Moments of the distribution.* We now derive a recurrence relation from which the moments of the distribution can be calculated. We denote

$$\int_0^\infty\left(\sum_{k=r}^{n-r}{}^{n-2r}C_{k-r}\,t^s\exp\left[-\frac{nt^2}{2k(n-k)}\right]G_{k-1}(t)\,G_{n-k-1}(t)\right)dt$$

by $I(n,r,s)$ $(s\geqslant 0)$. (The summation is from 1 to $n-1$ when $r = 0$.) For $s\geqslant 1$ we have, on integrating by parts, suitably grouping terms and, in one case, changing the summation variable from k to $k+1$, the successive equations

$$I(n,r,s) = \int_0^\infty\Sigma\,{}^{n-2r}C_{k-r}\frac{k(n-k)}{n}\exp\left[-\frac{nt^2}{2k(n-k)}\right]$$
$$\times\left\{(s-1)\,t^{s-2}G_{k-1}(t)\,G_{n-k-1}(t)+t^{s-1}\exp\left[-\frac{t^2}{2k(k-1)}\right]G_{k-2}(t)\,G_{n-k-1}(t)\right.$$
$$\left.+\,t^{s-1}\exp\left[-\frac{t^2}{2(n-k-1)(n-k)}\right]G_{k-1}(t)\,G_{n-k-2}(t)\right\}dt$$
$$= \int_0^\infty\Sigma\,{}^{n-2r}C_{k-r}\frac{r(n-r)+(k-r)(n-k-r)}{n}(s-1)\,t^{s-2}\exp\left[-\frac{nt^2}{2k(n-k)}\right]G_{k-1}(t)\,G_{n-k-1}(t)\,dt$$
$$+\int_0^\infty\Sigma\exp\left[-\frac{(n-1)\,t^2}{2k(n-k-1)}\right]G_{k-1}(t)\,G_{n-k-2}(t)\,t^{s-1}$$
$$\times\left[\frac{k(n-k)}{n}\,{}^{n-2r}C_{k-r}+\frac{(k+1)(n-k-1)}{n}\,{}^{n-2r}C_{k-r+1}\right]dt,$$

and, since
$$\frac{k(n-k)}{n}{}_{n-2r}C_{k-r}+\frac{(k+1)(n-k-1)}{n}{}_{n-2r}C_{k-r+1}$$
$$=\frac{(n-2r)(n-2r-1)}{n}{}_{n-2r-1}C_{k-r}+\frac{r(n-r)}{n}{}_{n-2r+1}C_{k-r+1},$$
we have
$$I(n,r,s)=\frac{r(n-r)(s-1)}{n}I(n,r,s-2)+\frac{(n-2r)(n-2r-1)(s-1)}{n}I(n,r+1,s-2)$$
$$+\frac{(n-2r)(n-2r-1)}{n}I(n-1,r,s-1)+\frac{r(n-r)}{n}I(n-1,r-1,s-1). \quad (4)$$

Now the sth moment of the distribution of m is $n^{\frac{1}{2}}(2\pi)^{-\frac{1}{2}(n-1)}(2/n)^s I(n,0,s)$ and by the aid of (4) we can express $I(n,0,s)$ in terms of certain of the $I(n,r,0)$, $I(n-1,r,0)$, etc. (For a given value of s, we need values of r for which $0\leqslant 2r\leqslant s$.) $I(n,r,0)$ is the integral from 0 to ∞ of certain terms in the frequency function of m, viz. ${}_{n-2r}C_{k-r}$ of those arising when k of the d's are negative ($r\leqslant k\leqslant n-r$). To get the whole frequency function ${}_nC_k$ terms are taken because that is the number of ways of choosing k negative d's. But if r of the d's are fixed as positive and r as negative (i.e. a certain part only of the sample space is considered), then the remaining $(k-r)$ negative d's can be chosen in ${}_{n-2r}C_{k-r}$ ways. Hence $I(n,r,0)$ is the integral of $(2\pi n)^{-\frac{1}{2}}\exp[-\frac{1}{2}\Sigma x^2]$ over this restricted region of the sample space. In particular, $I(n,0,0)=(2\pi)^{\frac{1}{2}(n-1)}n^{-\frac{1}{2}}$, a result which also follows from the frequency function of m.

To evaluate the integral it is convenient to transform the sample space by the matrix T, k being taken equal to 1. (This does not imply as much as before about the relation of the mean to the other observations, since we shall now only restrict the signs of two of the d's.) For $r=1$, this gives the definition of the restricted region to be
$$z_2>0, \quad \frac{z_2}{\sqrt{\{(n(n-1)\}}}+z_3\sqrt{\left(\frac{n-2}{n-1}\right)}>0.$$
The value of the integral of $(2\pi n)^{-\frac{1}{2}}\exp[-\frac{1}{2}\Sigma z^2]$ over this region bears to its value over all space the ratio (i) $\frac{1}{2}\pi+\sin^{-1}[1/(n-1)]$ to (ii) 2π, since (i) is the angle between the two hyperplanes bounding the region. Hence
$$\frac{I(n,1,0)}{I(n,0,0)}=\frac{1}{4}+\frac{1}{2\pi}\sin^{-1}\frac{1}{n-1}.$$

The calculation of $I(n,2,0)$ involves a much more complicated integration. The result is that
$$I(n,2,0)=\frac{(2\pi)^{\frac{1}{2}(n-1)}}{\sqrt{n}}\left[\frac{1}{16}+\frac{\sin^{-1}[1/(n-1)]+\sin^{-1}[1/(n-3)]}{8\pi}\right.$$
$$+\frac{5\{(\sin^{-1}[1/(n-1)])(\sin^{-1}[1/(n-3)])\}}{4\pi^2}$$
$$\left.-\frac{1}{\pi^2}\int_0^{\sin^{-1}[1/(n-3)]}\tan^{-1}\left(\frac{2-(n-1)(n-4)\tan^2\phi}{n(n-3)}\right)^{\frac{1}{2}}d\phi\right].$$

The result can be put in several forms, but this is the one which I personally have found most useful for computation.

For larger r the $I(n,r,0)$ would presumably be even more complicated; the two given are sufficient to determine the first five moments of the distribution. Since the d's tend to

independence for large n and the assignment of sign to a d reduces the sample space by one-half, we have that

$$I(n, r, 0) = \frac{(2\pi)^{\frac{1}{2}(n-1)}}{\sqrt{n}}\left(\frac{1}{4^r} + O\left(\frac{1}{n}\right)\right).$$

By the use of the recurrence relation (4) we can now find the moments and constants of shape of the distribution of m; these agree with the expressions obtained by Geary (1936) (see also some of his results quoted by Pearson (1945)), except that the coefficient of n^{-6} in the expansion of m_4'' (formula (22), 1936 paper) should be $(51 - 352a^2 + (2136/5)a^4)$ and not $(51 - 352a^2 + 427a^4)$. This mistake causes the coefficient of n^{-6} in the expansion of $m_4'' - m_4'$ (formula (24), same paper) to be 0·033578 instead of the correct 0·114635, and that of ν^{-6} in the expansion of λ_4 (formula (4), 1945) to be $-0{\cdot}120003$ instead of the correct $-0{\cdot}038946$.

We also have the new result

$$\mu_5 = \frac{0{\cdot}792218}{n^3} + \frac{0{\cdot}023893}{n^4} + \frac{0{\cdot}097967}{n^5} + \frac{0{\cdot}133384}{n^6} + \ldots.$$

5. *Approximations to the distribution.* The exact calculation of tables of the probability integral of m is lengthy, and the labour increases as n does. Consequently it seemed worth while to investigate the accuracy of approximations to the distribution which would involve less work in their computation. This investigation was of an empirical nature and consisted of comparing the values given for the percentage points of the distribution of m for $n = 10$ with the true values calculated by Hartley (1945). It was assumed that the approximation which is most accurate for that value of n will also be the most accurate for other values. The approximations considered were

(*a*) the normal distribution with the same first two moments as m;

(*b*) the distribution of the form $Km^\alpha e^{-\beta m^2}$ with the same first two moments as m;

(*c*) the Pearson curve (in fact, Type I) with the same first four moments as m;

(*d*) functions of m chosen so that their distributions are more nearly normal than that of m.

Of these methods, (*d*) is the most useful; however, a brief description of the others may be of interest.

(*a*), as may be seen by the comparison in Table 2 following my earlier paper, is useless for small n. Comparing it with (*d*) for $n = 1000$, it is found that the percentage points are given with an error of about 0·01 in the extreme values (0·1 and 99·9 %).

(*b*), when fitted to sample size 10 gives the 99 % point with error 0·01 and the 99·9 % point with error 0·02.

(*c*), gives percentage points for sample size 10 correct to three places of decimals for cumulative probabilities between 1 and 99·8 %. The computation is laborious, however, and gets heavier for large n, owing to the difficulty of calculating values of the incomplete B-function for large p, q.

(*d*) is an application of a method due to Haldane (1937). The distributions of the functions $y = (m/\overline{m})^k$ and $z = (1 + (m - \overline{m})/g)^h$ are considered, where k, g and h are chosen so that for the distribution of y, β_1 is $O(n^{-3})$, while for the distribution of z, β_1 is $O(n^{-3})$ and $\beta_2 - 3$ is $O(n^{-2})$. To find k, g and h, powers of y and z are expanded in powers of $(m - \overline{m})$, and their products with $f(m)\,dm$ integrated with m ranging from 0 to ∞. The moments of y and z are thus found in terms of the moments of m and can be arranged as power series in n^{-1}. (In a review of Haldane's paper, Neyman (1938) points out that this method is not rigorous as the series in $(m - \overline{m})$ are divergent for some values of m; this does not, however, necessarily imply that the expansions of the moments of y and z in terms of n^{-1}, as far as is needed, are

false. An investigation of the remainder terms of these series would be of interest.) k, g and h are now found so as to make the moments of y and z of the required order. To find g and h a knowledge of the sixth moment of m is needed. To avoid calculation of $I(n, 3, 0)$ it was assumed that, for the distribution of m,

$$\mu_6/(15\mu_2^3) = 1 + O(n^{-1});$$

this seems reasonable in view of the fact that $\mu_4/(3\mu_2^2)$, $\mu_5/(10\mu_2\mu_3)$ are both $1 + O(n^{-1})$. The results are that

$$y = (m/\overline{m})^{0\cdot5609}$$

is approximately normally distributed with mean

$$1 - \frac{0\cdot0703}{n} - \frac{0\cdot0714}{n^2} - \frac{0\cdot0818}{n^3} - \ldots,$$

and standard deviation

$$\frac{0\cdot42375}{\sqrt{n}}\left(1 + \frac{0\cdot4808}{n} + \frac{0\cdot4099}{n^2} + \ldots\right),$$

while

$$z = \left(1 + \frac{m - \overline{m}}{0\cdot6005}\right)^{0\cdot6695}$$

is approximately normally distributed with mean

$$1 - \frac{0\cdot1115}{n} - \frac{0\cdot0239}{n^2} - \frac{0\cdot0135}{n^3} - \ldots$$

and standard deviation

$$\frac{0\cdot67205}{\sqrt{n}}\left[1 + \frac{0\cdot0219}{n} + \frac{0\cdot0082}{n^2} + \ldots\right].$$

For sample size 10, y gives percentage points differing by amounts up to 0·006 (for 99 %) from the true values, while z gives values differing by not more than 0·001. The two methods give the same values (to three places of decimals) when n is greater than 35. The labour of calculation does not increase with n, and for a percentage point which is frequently needed, the equation giving y or z in terms of m can be inverted to give the desired value directly as a power series in $n^{-\frac{1}{2}}$.

REFERENCES

GEARY, R. C. (1936). *Biometrika*, **28**, 295–305.
GODWIN, H. J. (1945). *Biometrika*, **33**, 254–6.
HALDANE, J. B. S. (1937). *Biometrika*, **29**, 392–404.
HARTLEY, H. O. (1945). *Biometrika*, **33**, 257–65.
MCKAY, A. T. (1935). *Biometrika*, **27**, 466–71.
NEYMAN, J. (1938). *Zbl. Math.* **18**, 257.
PEARSON, E. S. (1945). *Biometrika*, **33**, 252–3 (Editorial Note).

A NOTE ON THE ASYMPTOTIC DISTRIBUTION OF RANGE

By D. R. COX, *Wool Industries Research Association*

1. Introduction and summary

In a recent paper Elfving (1947) has given an asymptotic form for the distribution of range in large samples.

In the present note, two other methods of obtaining an asymptotic form for the range distribution are discussed, which have the advantage of being expressed directly in terms of the range, while Elfving's form involved a non-linear transformation of range.

Numerical results are given for a normal population comparing the exact distribution of range with the two approximations discussed here, and with Elfving's approximation.

2. Derivation from Fisher and Tippett's results on the distribution of extremes

Fisher & Tippett (1928) have obtained results for the distribution of the least and greatest members of large random samples. Now in large samples it is clear that least and greatest values are effectively independent, and so a form for the asymptotic distribution of range can be derived by integration from the joint distribution of least and greatest values.*

Consider random samples of size n taken from a population whose frequency function is $\phi(x)$ and whose distribution function is $\Phi(x)$. Define $x_n^{(1)}$ and $x_n^{(2)}$ by

$$\Phi(x_n^{(1)}) = 1 - \frac{1}{n}, \tag{1}$$

$$\Phi(x_n^{(2)}) = \frac{1}{n}. \tag{2}$$

Let

$$y_1 = n(x_1 - x_n^{(1)})\,\phi(x_n^{(1)}), \tag{3}$$

$$y_2 = n(x_2 - x_n^{(2)})\,\phi(x_n^{(2)}), \tag{4}$$

where x_1 and x_2 are the greatest and least members of the sample.

Then Fisher & Tippett have shown that if $\phi(x)$ tends to zero exponentially or faster as x tends to infinity, the limiting frequency functions of y_1 and y_2 are

$$\exp(-y_1 - e^{-y_1}) \quad \text{and} \quad \exp(y_2 - e^{y_2}).$$

Thus the limiting joint-frequency function of y_1 and y_2 is

$$\exp(-y_1 - e^{-y_1} + y_2 - e^{y_2}),$$

and so if

$$W = y_1 - y_2, \tag{5}$$

the limiting frequency function of W is

$$\int_{-\infty}^{\infty} \exp(-W - e^{-W-y_2} - e^{y_2})\,dy_2 = 2e^{-W}K_0(2e^{-\frac{1}{2}W}). \tag{6}$$

In equation (6), $K_0(x)$ is a modified Bessel function of the second kind (Watson, 1944, p. 78).

* Since I wrote this paper, it has been pointed out to me that this method has recently been discussed at length by E. J. Gumbel (1947). In the case of a symmetrical distribution, W of equation (5) below is termed by Gumbel the 'reduced range' and denoted by R. For the case of a normal population he compares the exact probability levels of the sample range with those derived from the asymptotic distribution of equation (6) for which he has calculated certain values of the probability integral. He has not, however, considered the rather closer approximation which I discuss in § 3 below.

Suppose now that the basic distribution is symmetrical and has mean zero, so that

$$x_n^{(1)} = -x_n^{(2)},$$

and

$$\phi(x_n^{(1)}) = \phi(x_n^{(2)}).$$

Then if w_n is the sample range

$$w_n = 2x_n^{(2)} + \frac{W}{n\phi(x_n^{(2)})}. \tag{7}$$

Thus (6) leads immediately to a form for the asymptotic frequency function of range.

It can be shown directly from (6), or inferred from Fisher & Tippett's work, that the mean and standard deviation of W in the distribution (6), are

$$\overline{W} = 2\gamma, \tag{8}$$

$$\sigma_W = \frac{\pi}{\sqrt{3}}, \tag{9}$$

where γ is Euler's constant.

Also

$$\beta_1 = 0{\cdot}806, \tag{10}$$

and

$$\beta_2 = 4{\cdot}2. \tag{11}$$

If we combine these results with equation (7), expressing the range w_n in terms of W, we find that for large n

$$\overline{w}_n = 2x_n^{(2)} + \frac{2\gamma}{n\phi(x_n^{(2)})}, \tag{12}$$

$$\sigma_{w_n} = \frac{\pi}{\sqrt{3}\, n\phi(x_n^{(2)})}. \tag{13}$$

The limiting β coefficients of w_n are the same as those of W.

A numerical comparison of these results with the exact results for the normal distribution is given in § 4.

3. Asymptotic form by the method of steepest descents

The exact expression for the frequency function of range involves an integral. For large n, this integral can be evaluated by the method of steepest descents and this leads to a further form for the asymptotic frequency function of range.

The frequency function, $f_n(w_n)$, of the range w_n in samples of size n, is known to be

$$\begin{aligned} f_n(w_n) &= n(n-1)\int_{-\infty}^{\infty} \phi(u-\tfrac{1}{2}w_n)\,\phi(u+\tfrac{1}{2}w_n)\,[\Phi(u+\tfrac{1}{2}w_n)-\Phi(u-\tfrac{1}{2}w_n)]^{n-2}\,du \\ &= n(n-1)\int_{-\infty}^{\infty} \phi(u-\tfrac{1}{2}w_n)\,\phi(u+\tfrac{1}{2}w_n)\exp[(n-2)\,\psi(u,w_n)]\,du, \end{aligned} \tag{14}$$

where

$$\psi(u,w_n) = \log[\Phi(u+\tfrac{1}{2}w_n)-\Phi(u-\tfrac{1}{2}w_n)]. \tag{15}$$

It will now be supposed for simplicity that $\phi(x)$ is a symmetrical unimodal frequency distribution with mean zero; that is, that

$$\phi(x) = \phi(-x), \qquad \phi'(x) \neq 0 \quad (x \neq 0).$$

Then for fixed w_n, the function of u, $\psi(u,w_n)$ has a single maximum at $u = 0$. The integral (14) is thus of the form that can be evaluated by the method of steepest descents (Watson, 1944, p. 235).

The basic idea is that as n tends to infinity, by far the greatest portion of the integral (14) comes from the neighbourhood of $u = 0$.

Write
$$\psi(u, w_n) = \psi(0, w_n) - \tfrac{1}{2}t^2, \tag{16}$$
and then it is found that
$$f_n(w_n) = n(n-1)\exp[(n-2)\psi(0, w_n)]\int_{-\infty}^{\infty} \phi(u-\tfrac{1}{2}w_n)\phi(u+\tfrac{1}{2}w_n)\frac{du}{dt}\exp\left[-\frac{(n-2)t^2}{2}\right]dt. \tag{17}$$

The equation (16) can be solved for u as a power series in t, and so the expression
$$\phi(u-\tfrac{1}{2}w_n)\phi(u+\tfrac{1}{2}w_n)\frac{du}{dt}$$
can be obtained as a function of t.

It is found that
$$\phi(u-\tfrac{1}{2}w_n)\phi(u+\tfrac{1}{2}w_n)\frac{du}{dt} = \frac{[\phi(\tfrac{1}{2}w_n)]^2}{[-\psi''(0, w_n)]^{\frac{1}{2}}}[1 + A_2(w_n)t^2 + \ldots].$$

In this expression $A_2(w_n)$ is a complicated function of $\Phi(\frac{1}{2}w_n)$, $\phi(\frac{1}{2}w_n)$, ..., $\phi''(\frac{1}{2}w_n)$, whose exact form will not be needed, and
$$\psi''(0, w_n) = \left.\frac{\partial^2\psi(u, w_n)}{\partial u^2}\right|_{u=0}.$$

Thus
$$f_n(w_n) = \frac{n(n-1)\exp[(n-2)\psi(0, w_n)][\phi(\tfrac{1}{2}w_n)]^2}{[-\psi''(0, w_n)]^{\frac{1}{2}}}\int_{-\infty}^{\infty}[1 + A_2(w_n)t^2 + \ldots]\exp\left[-\frac{(n-2)t^2}{2}\right]dt. \tag{18}$$

An asymptotic expansion in inverse powers of $(n-2)$ follows if the integral may be evaluated formally term by term. That this is in fact permissible may be shown by an application of Watson's lemma (Watson, 1944, p. 236).

Thus
$$f_n(w_n) \sim \frac{n(n-1)(2\pi)^{\frac{1}{2}}\exp[(n-2)\psi(0, w_n)][\phi(\tfrac{1}{2}w_n)]^2}{[-\psi''(0, w_n)]^{\frac{1}{2}}}\left\{\frac{1}{(n-2)^{\frac{1}{2}}} + \frac{A_2(w_n)}{(n-2)^{\frac{1}{2}}} + \ldots\right\}$$
$$\sim \frac{n^{\frac{1}{2}}(2\pi)^{\frac{1}{2}}[\phi(\tfrac{1}{2}w_n)]^2}{[\phi'(-\tfrac{1}{2}w_n)]^{\frac{1}{2}}}[\Phi(\tfrac{1}{2}w_n) - \Phi(-\tfrac{1}{2}w_n)]^{n-\frac{1}{2}}\left\{1 + \frac{A_2(w_n)}{n-2} + \ldots\right\}. \tag{19}$$

To obtain the last expression, the definition (15) of $\psi(u, w_n)$ has been used.

Put
$$\Psi(x) = \Phi(x) - \Phi(-x). \tag{20}$$

Then the first term of the expansion (19) is
$$\frac{n^{\frac{1}{2}}(2\pi)^{\frac{1}{2}}[\phi(\tfrac{1}{2}w_n)]^2}{[\phi'(-\tfrac{1}{2}w_n)]^{\frac{1}{2}}}[\Psi(\tfrac{1}{2}w_n)]^{n-\frac{1}{2}}. \tag{21}$$

Now it can be shown that the function $A_2(w_n)$ is negative for the normal and many other distributions. A better approximation than (21) is therefore
$$\frac{c_n[\phi(\tfrac{1}{2}w_n)]^2[\Psi(\tfrac{1}{2}w_n)]^{n-\frac{1}{2}}}{[\phi'(-\tfrac{1}{2}w_n)]^{\frac{1}{2}}}, \tag{22}$$
where in this expression the constant c_n is to be chosen to make the integral from zero to infinity equal to unity.

In § 4 some numerical results are given comparing (22) with the exact distribution, for the case when the basic distribution is normal.

4. Numerical comparison with exact values

In this section the approximate results obtained above are compared with exact values for the normal distribution.

Figs. 1–4 give the frequency functions of range for $n = 20$ and $n = 50$ when the basic distribution is normal with unit standard deviation. The exact distribution was obtained from E. S. Pearson & H. O. Hartley's table (1942) for $n = 20$, and by numerical integration for $n = 50$.

Table 1 gives the exact values of mean, standard deviation, β_1 and β_2 of range, compared with the values from the formulae (8)–(11) of § 2, and with numerically computed values from the approximation of § 3 and from Elfving's approximation mentioned in § 1. The exact values, based on the work of Tippett (1925) and E. S. Pearson (1926), have been taken from *Tables for Statisticians and Biometricians*, Part II (K. Pearson, 1931, p. cxvii).

Table 1. *Constants of the range distribution in samples from a normal distribution of unit standard deviation*

n	Exact values				Approx. of § 2			
	Mean	S.D.	β_1	β_2	Mean	S.D.	β_1	β_2
10	3·08	0·80	0·16	3·22	3·22	1·03	0·81	4·2
20	3·73	0·73	0·16	3·26	3·85	0·88	0·81	4·2
50	4·50	0·65	0·19	3·34	4·58	0·75	0·81	4·2
100	5·02	0·60	0·22	3·39	5·09	0·68	0·81	4·2
500	6·07	0·52	0·28	3·50	6·12	0·57	0·81	4·2
1000	6·48	0·50	0·31	3·54	6·52	0·54	0·81	4·2

n	Approx. of § 3				Elfving's approx.			
	Mean	S.D.	β_1	β_2	Mean	S.D.	β_1	β_2
20	3·83	0·78	0·21	3·42	3·76	0·77	0·07	3·08
50	4·58	0·69	0·24	3·33	4·52	0·67	0·13	3·18

Figs. 1–4 show that the steepest descents approximation is more accurate than the approximation of § 2, although even for $n = 20$ the latter gives the ordinates of the range distribution fairly closely in the centre of the distribution but not in the tails.

This is confirmed by the moments given in the table. The Bessel function approximation of § 2 gives the mean range reasonably accurately, the standard deviation less accurately, while the β coefficients are very different (for the range of values of n considered) from the exact values. This is in accordance with the results of Fisher & Tippett, who showed that for the distribution of extremes the limiting β values are not reached until n is about 10^{12}.

The steepest descents approximation of § 3, and Elfving's approximation mentioned in § 1 are both more accurate. For $n = 50$ Elfving's approximation gives the mean range to

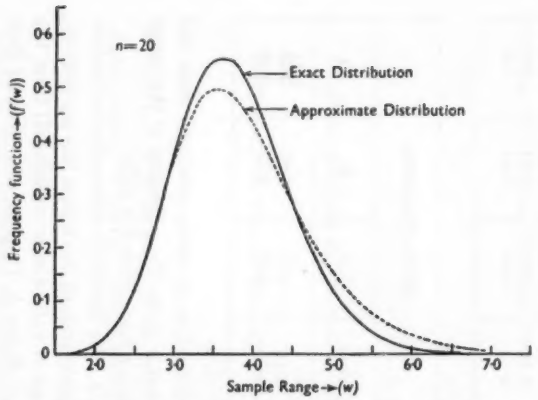

Fig. 1. Exact form compared with Bessel function approximation of § 2.

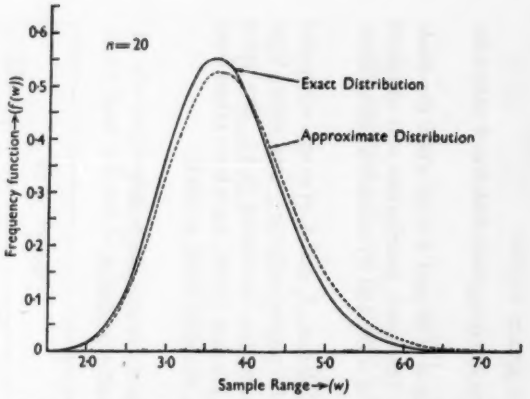

Fig. 2. Exact form compared with steepest descents approximation of § 3.

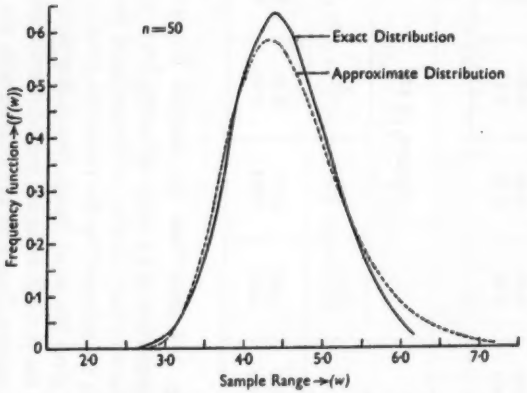

Fig. 3. Exact form compared with Bessel function approximation of § 2.

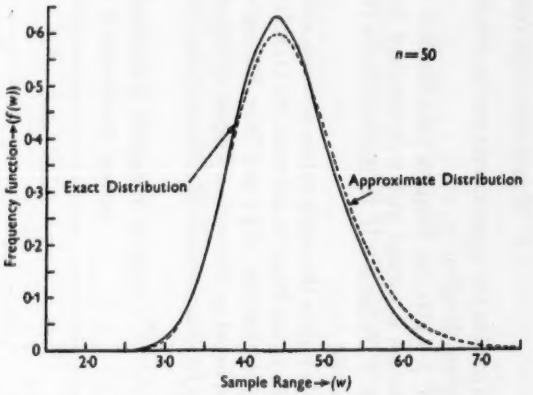

Fig. 4. Exact form compared with steepest descents approximation of § 3.

Fig. 1–4. Distribution of range in samples from normal population of unit standard deviation.

$\frac{1}{2}$ % and the standard deviation of range to 3 %. The steepest descents approximation is less accurate, the corresponding figures being $1\frac{1}{2}$ and 6 %. The β coefficients appear to be given slightly more accurately by the steepest descents approximation.

The properties of the three approximations may be summarized as follows for the case when the basic distribution is normal.

Elfving's approximation is the most accurate with the steepest descents approximation second.

The Bessel function approximation gives fairly accurate estimates of the ordinates of the range distribution near the mode, but does not reproduce the ordinates in the tails until n is very large indeed.

The disadvantage of Elfving's method is that it involves a non-linear transformation of the range. This makes it unlikely that it can be used algebraically to simplify mathematical problems involving the distribution of range.

I wish to thank the Director of Research, Wool Industries Research Association, for permission to publish this paper, and Dr H. E. Daniels for his interest in the work.

REFERENCES

Elfving, G. (1947). *Biometrika*, **34**, 111.
Fisher, R. A. & Tippett, L. H. C. (1928). *Proc. Camb. Phil. Soc.* **24**, 180.
Gumbel, E. J. (1947). *Ann. Math. Statist.* **18**, 384.
Pearson, E. S. (1926). *Biometrika*, **18**, 173.
Pearson, E. S. & Hartley, H. O. (1942). *Biometrika*, **33**, 301.
Pearson, K. (1931). *Tables for Statisticians and Biometricians*, Part II. Cambridge University Press.
Tippett, L. H. C. (1925). *Biometrika*, **17**, 364.
Watson, G. N. (1944). *Theory of Bessel Functions*, 2nd ed. Cambridge University Press.

Fig. 4. Exact form compared with steepest descents approximation of § 3.

Fig. 3. Exact form compared with Bessel function approximation of § 2.

Fig. 1–4. Distribution of range in samples from normal population of unit standard deviation.

ON THE ROLE OF VARIABLE GENERATION TIME IN THE DEVELOPMENT OF A STOCHASTIC BIRTH PROCESS

By DAVID G. KENDALL, *Magdalen College, Oxford*

1. **Introduction.** The 'birth-and-death' process introduced by W. Feller (1939) provides a mathematical description of the growth of a population under the influence of very much simplified laws of reproduction and mortality. The population size n is considered as a random function of the time t, and the equations governing its development are derived from the following assumptions:

(i) The value of n at some initial time ($t = 0$) is supposed given. In the simplest case $n(0) = 1$.

(ii) It is assumed that the subpopulations stemming from two co-existing individuals will develop in complete independence of one another.

(iii) The risks of mortality and reproduction are supposed to be the same for each member of the population.

(iv) An individual known to be alive at time t has a chance asymptotically equal to λdt of reproducing itself, and a chance asymptotically equal to μdt of dying during the subsequent elementary time interval of length dt. Reproduction is here taken to imply binary subdivision, and so to result in the addition of just one member to the population.

(v) The chances of reproduction and mortality, described at (iv) above, are supposed to be completely independent of the previous history of the individual, and in particular to be independent of the time which has elapsed since its own 'birth' in some earlier subdivision.

(vi) The birth and death rates λ and μ are supposed to be independent of the epoch t.

Assumption (v) implies that the stochastic dependence of n upon t can be described with the aid of a discontinuous Markoff process,* and, in fact, it is easily seen that the functions

$$P_n(t) \equiv \text{Probability}\{n(t) = n \mid n(0) = 1\},$$

which completely determine the structure of the process, must satisfy the set of differential-difference equations

$$\left.\begin{aligned} \frac{d}{dt}P_n(t) &= (n+1)\mu P_{n+1}(t) + (n-1)\lambda P_{n-1}(t) - n(\lambda+\mu)P_n(t) \quad (n \geqslant 1), \\ \frac{d}{dt}P_0(t) &= \mu P_1(t). \end{aligned}\right\} \tag{1}$$

Feller's equations (1) were solved by C. Palm,† the mean and variance of n as functions of t having been already given in Feller's paper.

The idealized population whose growth is described by the equations (1) is, of course, rather far removed from reality, and it is therefore of some interest to consider how far any

* The development of a system is said to follow a Markoff process if its state at any time t can be described by the value of a random time-dependent variable $X(t)$ with the following property: *let the value of $X(t_0)$ be known; then if $t_1 > t_0$, the conditional distribution of the random variable $X(t_1)$ is in no way affected if the value of $X(t)$ is also given for any $t < t_0$.* Further details and some references will be found in my recent review (Kendall, 1947).

† Palm's formulae are quoted by N. Arley & V. Borchsenius (1945). See also M. S. Bartlett (1947) and D. G. Kendall (1948*a*).

relaxation is possible in the assumptions (i)–(vi). Of these the first two and the last two are the ones urgently requiring consideration; progress is most likely to be made by attacking them independently, and a good deal of work has already been done in this respect. If (ii) is retained, (i) is easily relaxed, for the $n(0)$ initial individuals then generate independent subpopulations each commencing with a single member, and it is therefore sufficient to raise the generating function for the $\{P_n(t)\}$ to the power $n(0)$. An interesting extension of (ii) has been considered by Feller, who discusses the birth-and-death process for which the rates λ and μ, instead of being constants, are linearly dependent on the instantaneous population size n. This model, the equations for which are as yet unsolved, corresponds to a population growing according to the logistic law of Pearl, Verhulst and Reed in the parallel deterministic theory. In a recent paper (Kendall, 1948*b*) I have given an account of the birth-and-death process in which the birth- and death-rates λ and μ, instead of being constants as in (vi), can be any desired functions of the epoch t; the most interesting example is perhaps that in which λ and μ are periodic functions of t.

The one important assumption which has not yet been relaxed is that of the Markoff property (v). This implies, for example, that in the absence of mortality an individual 'born' at time t will itself undergo subdivision at a time $t+\tau$, where the *generation time* τ has the distribution

$$e^{-\lambda\tau}\lambda\,d\tau \quad (0<\tau<\infty). \tag{2}$$

This is, of course, very different from the distributions of generation time actually observed; not only for man, but also for such elementary organisms as bacteria, the observed distribution usually possesses a pronounced non-zero mode, and the other extreme assumption of a *fixed* generation time $\tau=\tau_0$ (implying an exact doubling of the population at regular intervals) might seem to be more realistic.

Now the distribution (2) would assert that τ is a multiple of a χ^2-variate having two degrees of freedom, and this suggests that it would be worth while examining a modified process in which τ is distributed as the multiple of a χ^2_{2k}, where k is an integer greater than unity. The present paper is devoted to a development of this idea, and to simplify the analysis attention will here be confined to purely reproductive processes ($\mu=0$). Before proceeding to details it is of interest to note what values of k are likely to be of practical relevance.

The most extensive information on the distribution of generation times for bacteria appears to be that communicated by C. D. Kelly & Otto Rahn (1932). They maintained bacteria in a 'warm stage' at 30° C., and measured (by continuous microscopic observation) the fission times for the second, third and fourth generations descended from each of a large number of individual cells. Their results concerning *Bacterium aerogenes* may be quoted in illustration. Observations were made, for this bacterium, on nine different days, the number of generation times measured per day varying from 30 to 126; the results for the three largest sets (each consisting of more than a hundred observations) are shown in histogram form in the accompanying diagram (Fig. 1). The sets relating to the several days have not been pooled because it is evident that there was a day-to-day variation in the mean generation time.

It will be seen at once that the assumption of a χ^2_{2k} distribution (with a change of scale to allow for the correct mean time) is at least not an unreasonable one, and that it is greatly superior to either of the alternatives previously available (which correspond to $k=1$ and $k=\infty$). While it will certainly be of interest to examine the adequacy of the χ^2_{2k} hypothesis in more detail on another occasion, it is enough for the moment to know that it provides

a way of introducing into a model of population growth a law of variation of generation time which bears a general resemblance to that obtaining in reality.

It will be useful for the further developments of this paper to have in mind a rough estimate of the parameter k derived from the above data. Such an estimate can conveniently be based

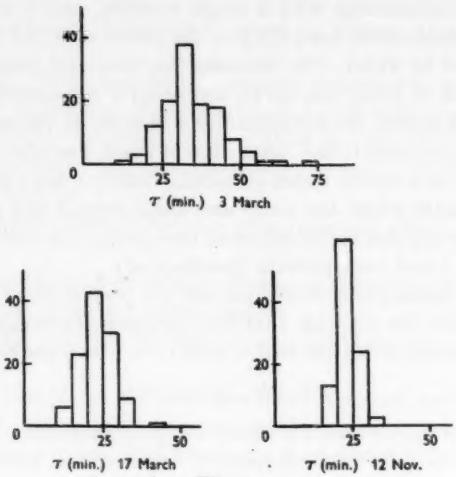

Fig. 1

Table 1. Bacterium aerogenes *at* 30° C. *Preliminary analysis of the observed generation times*, τ

(Data of Kelly and Rahn)

Date	No. of observations	Geometric mean of τ (min.)	'Coefficient of variation' (%)	Estimate of k
17 Feb.	44	38·5	17·6	33
24 Feb.	60	34·8	22·8	20
2 Mar.	84	28·9	27·7	14
3 Mar.	126	33·5	27·2	14
6 Mar.	84	29·6	23·0	19
10 Mar.	93	32·4	30·1	12
17 Mar.	112	22·6	23·9	18
12 Nov.	100	23·1	14·6	47
14 Nov.	30	24·9	24·7	17

on the observed variance of the natural logarithm of the generation time, which on the χ^2_{2k} hypothesis will have the expected value*

$$\kappa_2(\log \tau) = 1/(k-\tfrac{1}{2}), \quad \text{approximately.}$$

It is worth noting that the observed standard deviation of the logarithm of the generation time τ is roughly the same thing as the coefficient of variation of τ—it has been listed as such in Table 1. The observed variances of $\log \tau$, when analysed by M. S. Bartlett's well-known

* See, for example, M. S. Bartlett & D. G. Kendall (1946). (The error involved in this approximation is less than 1% for $k > 3$.)

test for homogeneity, were found to be significantly discordant ($\chi^2_8 = 62$), and no attempt has been made, therefore, to pool the nine estimates of k. For present purposes only the order of magnitude of k is required, and the value

$$k = 20$$

is evidently satisfactory, though, of course, there is no reason to suppose that it will be appropriate to describe the growth of any organism other than the one example discussed here.

2. **Specification of the multiple-phase birth process. The fundamental equations.** The process to be considered may be specified as follows. When a new individual is 'born', it passes through a series of phases, k in number, and only after it has attained the kth phase can it undergo subdivision. The lifetime in each phase is assumed to follow the law of distribution

$$e^{-k\lambda T} k\lambda dT \quad (0 < T < \infty),$$

the several lifetimes being independent, and the incident terminating the life of the individual in the kth phase being (i) its death and (ii) the birth, simultaneously, of two individuals who commence their existence in the first phase at that instant. Of course, the formal treatment would not in any way be altered if instead it were assumed that an individual terminates its existence in the kth phase by giving birth to one individual in the first phase, and simultaneously returning to the latter itself.

The multiple-phase birth process possesses the Markoff property, provided that its development is described by the *vector* variate $\mathbf{n}$, the k components of which enumerate the individuals existing in each of the k phases. If $\mathbf{n} \equiv (n_1, n_2, \ldots, n_k)$, and

$$n = n_1 + n_2 + \ldots + n_k,$$

then n, the total population size irrespective of phase, describes the growth of a birth process in which the generation time τ has the distribution of

$$\frac{1}{2k\lambda} \chi^2_{2k}.$$

When $k = 1$, the multiple-phase process is identical with the simple birth process discussed by Feller (i.e. that governed by equations (1), with $\mu = 0$), while if k is allowed to tend to infinity, one obtains the deterministic birth process with a fixed generation time $\tau_0 = 1/\lambda$. The multiple-phase process thus includes both the models already mentioned, and bridges the gap between them. Incidentally its study may throw some light on the effect of the 'Markoff' assumption in calculations of this sort. In this context, reference may be made to the general remarks in my review article (1947).

It will be supposed that initially the population consists of a single individual in the first phase. (Because of a well-known property of the distribution (2), it makes no difference whether this individual was born at the time $t = 0$ or at some earlier instant.) The course of events leading up to the first subdivision in what might be a typical case is shown in the accompanying diagram (Fig. 2). The stochastic development of the process will be fully described once the function

$$P\begin{pmatrix} n_k \\ \vdots \\ n_2 \\ n_1 \end{pmatrix}, \quad t \Bigg).$$

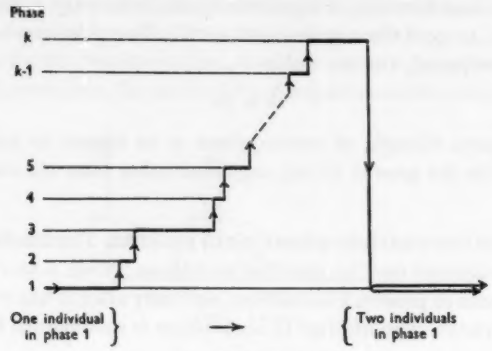

Fig. 2

is known; this is, the probability that at time t there will be n_1 individuals in the first phase, n_2 in the second phase, and so on. The differential-difference equations which correspond here to Feller's equations (1) can be written most concisely if one adopts the convention that $P \equiv 0$ whenever any of the n_i are negative. It will then readily be seen that

$$\frac{d}{dt} P\begin{pmatrix} n_k \\ \vdots \\ n_2 \\ n_1 \end{pmatrix}, \quad t \Bigg) = (n_1+1)\, k\lambda P\begin{pmatrix} n_k \\ \vdots \\ n_2-1 \\ n_1+1 \end{pmatrix}, \quad t \Bigg) + \dots$$

$$+ (n_{k-1}+1)\, k\lambda P\begin{pmatrix} n_k-1 \\ n_{k-1}+1 \\ \vdots \\ n_1 \end{pmatrix}, \quad t \Bigg) + (n_k+1)\, k\lambda P\begin{pmatrix} n_k+1 \\ \vdots \\ n_2 \\ n_1-2 \end{pmatrix}, \quad t \Bigg)$$

$$- (n_1+n_2+\dots+n_k)\, k\lambda P\begin{pmatrix} n_k \\ \vdots \\ n_2 \\ n_1 \end{pmatrix}, \quad t \Bigg), \tag{3}$$

and if one introduces the generating function

$$\phi\begin{pmatrix} z_k \\ \vdots \\ z_2 \\ z_1 \end{pmatrix}, \quad t \Bigg) = \Sigma z_1^{n_1} z_2^{n_2} \dots z_k^{n_k} P\begin{pmatrix} n_k \\ \vdots \\ n_2 \\ n_1 \end{pmatrix}, \quad t \Bigg), \tag{4}$$

it will follow that this must satisfy the partial differential equation

$$\frac{1}{k\lambda}\frac{\partial \phi}{\partial t} = \left(z_2 \frac{\partial \phi}{\partial z_1} + z_3 \frac{\partial \phi}{\partial z_2} + \dots + z_k \frac{\partial \phi}{\partial z_{k-1}}\right) + z_1^2 \frac{\partial \phi}{\partial z_k} - \sum_{i=1}^{k} z_i \frac{\partial \phi}{\partial z_i}, \tag{5}$$

the associated boundary condition being of course

$$\phi \equiv z_1 \quad \text{when} \quad t = 0. \tag{6}$$

The partial differential equation (5) is of the standard Lagrangian form, the auxiliary equations being

$$\frac{d\phi}{0} = \frac{k\lambda\, dt}{1} = \frac{dz_1}{z_1 - z_2} = \frac{dz_2}{z_2 - z_3} = \dots = \frac{dz_{k-1}}{z_{k-1} - z_k} = \frac{dz_k}{z_k - z_1^2}. \tag{7}$$

To solve these, it is convenient to introduce a new time scale defined by $k\lambda t = \theta$, and to write $X_i = e^{-\theta} z_i$; the equations (7) then become

$$X_j' = -X_{j+1} \quad (j = 1, 2, \ldots, k-1), \qquad X_k' = -e^{\theta} X_1^2,$$

and so

$$X_i = \left(-\frac{d}{d\theta}\right)^{i-1} X(\theta), \tag{8}$$

where $X(x)$ satisfies the ordinary differential equation

$$\left(-\frac{d}{dx}\right)^k X(x) = e^x \{X(x)\}^2. \tag{9}$$

Let the general solution of (9) be

$$X(x) = F(x; c_1, c_2, \ldots, c_k),$$

where the arbitrary constants $\{c_i\}$ are so chosen that

$$c_i = F^{(i-1)}(0; c_1, c_2, \ldots, c_k) \quad (1 \leqslant i \leqslant k). \tag{10}$$

Then

$$X_i = (-)^{i-1} F^{(i-1)}(\theta; c_1, c_2, \ldots, c_k), \tag{11}$$

and if these equations when solved for $\{c_j\}$ take the form

$$c_j = G_j(\theta; X_1, X_2, \ldots, X_k), \tag{12}$$

the general solution to (5) will be*

$$\phi(z_1, z_2, \ldots, z_k; t) \equiv \Phi(G_1, G_2, \ldots, G_k), \tag{13}$$

where Φ is an arbitrary function of its k arguments.

The unknown function Φ must now be identified with the aid of the boundary condition (6). To this end, put $\theta = 0$ in (11) and (12) and compare with (10); it will then be seen that

$$G_j(0; u_1, u_2, \ldots, u_k) \equiv (-)^{j-1} u_j.$$

From the boundary condition (6), however, it follows that

$$\begin{aligned} z_1 &\equiv \phi(z_1, z_2, \ldots, z_k; 0) \equiv \Phi\{\ldots, G_j(0; z_1, z_2, \ldots, z_k), \ldots\} \\ &\equiv \Phi(z_1, -z_2, \ldots, (-)^{k-1} z_k), \end{aligned}$$

and so the function Φ must be given by

$$\Phi(G_1, G_2, \ldots, G_k) \equiv G_1.$$

The generating function ϕ is thus $G_1(\theta; X_1, X_2, \ldots, X_k)$, and this is the value of c_1, i.e. of $F(0; c_1, c_2, \ldots, c_k)$, when the $\{c_j\}$ are to be determined from the k equations

$$z_i e^{-\theta} = (-1)^{i-1} F^{(i-1)}(\theta; c_1, c_2, \ldots, c_k).$$

Expressing this result in different words: $\phi(z_1, z_2, \ldots, z_k; t)$ is equal to the value of $X(0)$, when $X(x)$ satisfies the equation (9) together with the boundary conditions

$$X(\theta) = z_1 e^{-\theta}, \quad X'(\theta) = -z_2 e^{-\theta}, \quad \ldots, \quad X^{(k-1)}(\theta) = (-)^{k-1} z_k e^{-\theta}.$$

Before working out an example in detail, it is convenient to throw the solution into a more familiar form by writing $x = \theta - u$ and $X(x) = e^{-\theta} Z(u)$. It will then be seen that *the generating function ϕ for the multiple-phase birth process is equal to $e^{-k\lambda t} Z(k\lambda t)$, where the function $Z(u)$ is determined by the differential equation*

$$\left(\frac{d}{du}\right)^k Z(u) = e^{-u} \{Z(u)\}^2, \tag{14}$$

and the boundary conditions

$$Z^{(i)}(0) = z_{i+1} \quad (0 \leqslant i \leqslant k-1). \tag{15}$$

* The arguments of ϕ will henceforth, for typographical convenience, be written horizontally.

This is a perfectly straightforward one-point boundary problem, although unfortunately the equation (14) seems to be intractable for values of k greater than unity. A determination of the P-functions for the multiple-phase process is therefore not practicable; it is, however, possible to discuss the mean and variance of the distribution, and its approximate normality for large k, without solving the equation (14). These matters will be considered in the following sections of the paper; for the moment it is of interest to leave the general argument and examine the solution of the equation (14) in the elementary case when $k = 1$.

The equation is then $Z' = e^{-u}Z^2$, with the single boundary condition $Z(0) = z$. The general solution is

$$\frac{1}{Z} - e^{-u} = C,$$

and on identifying the arbitrary constant this becomes

$$Z(u) = \frac{z}{1 - z(1 - e^{-u})}.$$

Accordingly the generating function is

$$\phi = \frac{ze^{-\lambda t}}{1 - z(1 - e^{-\lambda t})}$$

and

$$P(n, t) = e^{-\lambda t}(1 - e^{-\lambda t})^{n-1} \quad (n \geqslant 1),$$

in agreement with Feller's original calculation (1939, equation (17) with $N = 1$). For future reference it will be convenient to note here the mean and variance of this distribution, as already given by Feller; they are

$$E(n) = e^{\lambda t} \quad \text{and} \quad \mathrm{Var}\,(n) = e^{\lambda t}(e^{\lambda t} - 1). \tag{16}$$

3. **The mean growth of the process.** It is possible to deduce the differential equations satisfied by the expected values of the $\{n_j\}$ from the fundamental equation (14), but it is easier to derive them independently; in fact, simple considerations of continuity give at once

$$\frac{d\nu_1}{d\theta} = 2\nu_k - \nu_1,$$

and

$$\frac{d\nu_j}{d\theta} = \nu_{j-1} - \nu_j \quad (1 < j \leqslant k), \tag{17}$$

where ν_j is the expected value of n_j. If now one writes ρ_j for $\nu_j e^{\theta}$ these equations can readily be seen to be equivalent to

$$\rho_j = \left(\frac{d}{d\theta}\right)^{k-j} \rho,$$

where

$$\left(\frac{d}{d\theta}\right)^{k} \rho = 2\rho.$$

Thus if ω is the primitive kth root of unity, $\exp(2\pi i/k)$, then

$$\rho = \sum_{r=0}^{k-1} A_r \exp(2^{1/k}\omega^r\theta)$$

(where the quantities A_r are as yet undetermined constants), and

$$\rho_j = 2^{1-j/k} \sum_{r=0}^{k-1} A_r \omega^{-jr} \exp(2^{1/k}\omega^r\theta).$$

Now initially, when both t and θ are zero, $\rho_1 = 1$ and $\rho_j = 0$ $(1 < j \leqslant k)$; the $\{A_r\}$ must therefore have the values

$$A_r = \frac{1}{2k} 2^{1/k}\omega^r.$$

Accordingly, the expected number of individuals in the jth phase at time t is

$$E(n_j) = \frac{1}{k} 2^{(1-j)/k} \sum_{r=0}^{k-1} \omega^{-(j-1)r} \exp\{(2^{1/k}\omega^r - 1)\, k\lambda t\}, \tag{18}$$

while the expected total population size at the same time is

$$E(n) = \frac{1}{2k} \sum_{r=0}^{k-1} \frac{2^{1/k}\omega^r}{2^{1/k}\omega^r - 1} \exp\{(2^{1/k}\omega^r - 1)\, k\lambda t\}, \tag{19}$$

$$= e^{-k\lambda t} \sum_{m=0}^{\infty} 2^m \left\{ \frac{(k\lambda t)^{mk}}{(mk)!} + \frac{(k\lambda t)^{mk+1}}{(mk+1)!} + \ldots + \frac{(k\lambda t)^{mk+k-1}}{(mk+k-1)!} \right\}. \tag{20}$$

In the last formula, the jth term in each bracket is the contribution from the jth phase.

These formulae possess two limiting forms of practical interest. The first describes (for fixed k) the behaviour of the mean population size for large values of the time t. The other limiting form is obtained by fixing t and allowing k to approach infinity (so that the process approximates to the deterministic model for which the generation time has the fixed value $1/\lambda$). Care is required in interpreting the results because the two limiting processes do *not* in general commute.

For large values of t (k remaining fixed) the dominant term in (18) is always the first. Indeed, since

$$2^{1/k} \cos\frac{2\pi}{k} - 1 < 0 \quad \text{when} \quad 2 \leqslant k \leqslant 28,$$

the remaining terms represent damped oscillations for these values of k, while for larger values of k they do not themselves tend to zero but are still of a lower order than the first term. Since

$$\lim_{k\to\infty} k(x^{1/k} - 1) = \log x,$$

it is convenient to write

$$\alpha_k = k(2^{1/k} - 1) \quad (\to \log 2 = 0{\cdot}693, \text{ as } k \to \infty).$$

The expected population sizes for large t are then given asymptotically by

$$E(n_j) \simeq \frac{2^{-(j-1)/k}}{k} e^{\alpha_k \lambda t}, \tag{21}$$

and

$$E(n) \simeq \frac{2^{1/k}}{2\alpha_k} e^{\alpha_k \lambda t}. \tag{22}$$

It is rather curious that the coefficient of $e^{\alpha_k \lambda t}$ in (22) has the limit 0·721 as k approaches infinity, and not unity as one might have expected. This is an example of the non-commuting character of the two limiting processes.

The dependence of α_k upon k is shown by the following short table. As far as the ultimate rate of growth of the mean population size is concerned, the variability of the generation time evidently ceases to have much effect after k exceeds a value of about 35 (this is the value of k for which α_k differs from $\log 2$ by 1 %):

k =	1	2	3	4	5	10	15	20	25	30	∞
α_k =	1·000	0·828	0·780	0·757	0·744	0·718	0·709	0·705	0·703	0·701	0·693

It is more difficult to discuss the behaviour of $E(n)$ when k tends to infinity and t remains fixed. From (20) it follows that $E(n)$ is equal to the expected value of $2^{[N/k]}$, when N is a Poisson variable of mean value $k\lambda t$, and $[x]$ is used to denote the integer part of x. It can be shown from this that

$$\lim_{k\to\infty} E(n) = 2^{[\lambda t]} \tag{23}$$

when λt is not an integer;

while $$\lim_{k\to\infty} E(n) = \tfrac{3}{4}.2^{\lambda t} \qquad (23a)$$

when λt is an integer. The limit of $E(n)$ as k approaches infinity is thus a discontinuous function of the time t, whose value at a point of discontinuity is equal to the arithmetic mean of the left- and right-hand limits. The details of this calculation, though elementary, are a little tiresome, and are perhaps not worth reproducing here. Formula (23) of course expresses the fact that in the limit as k tends to infinity the multiple-phase model coincides with the deterministic model already mentioned.

It will be noticed from (21) that for fixed k and large values of t, the expected numbers of individuals in each of the several phases will be proportional to

$$1, \quad 2^{-1/k}, \quad 2^{-2/k}, \quad \ldots, \quad 2^{-(k-1)/k}. \qquad (24)$$

When $k = 2$, it is possible to express the formulae for the expected population sizes in closed form:

$$E(n_1) = e^{-2\lambda t}\cosh(\lambda t\, 2\sqrt{2}) \quad \text{and} \quad E(n_2) = \frac{1}{\sqrt{2}} e^{-2\lambda t}\sinh(\lambda t\, 2\sqrt{2}). \qquad (25)$$

4. **The differential equation for the cumulant-generating function.** The cumulant-generating function for the total population size is

$$\log E(e^{n\alpha}) = K(\alpha,\theta) = \log\phi(e^{\alpha}, e^{\alpha}, \ldots, e^{\alpha}; t) = -\theta + \log Z(\theta),$$

where the function $Z(\theta)$ is to be determined from the differential equation

$$\left(\frac{d}{d\theta}\right)^k Z(\theta) = e^{-\theta}\{Z(\theta)\}^2,$$

with the boundary conditions $Z^{(i)}(0) = e^{\alpha} \quad (0 \leqslant i \leqslant k-1)$.

If in these equations one writes $Z = e^{\theta+K}$, the boundary conditions become

$$\left(1+\frac{\partial}{\partial\theta}\right)^i e^{K(\alpha,\theta)} = e^{\alpha}, \quad \text{when} \quad \theta = 0 \quad (0 \leqslant i \leqslant k-1),$$

or

$$e^{K(\alpha,0)} = e^{\alpha} \quad \text{and} \quad \left(\frac{\partial}{\partial\theta}\right)^i e^{K(\alpha,0)} = 0 \quad (1 \leqslant i \leqslant k-1),$$

from which it is easily seen that

$$K(\alpha,0) = \alpha \quad \text{and} \quad \left(\frac{\partial}{\partial\theta}\right)^i K(\alpha,0) = 0 \quad (1 \leqslant i \leqslant k-1), \qquad (26)$$

while the differential equation for $K(\alpha,\theta)$ which is to be associated with these boundary conditions is

$$\left(1+\frac{\partial}{\partial\theta}\right)^k e^{K(\alpha,\theta)} = e^{2K(\alpha,\theta)}. \qquad (27)$$

For some purposes it is convenient to write $\theta = kT$, so that $T = \lambda t$, and then

$$\left(1+\frac{1}{k}\frac{\partial}{\partial T}\right)^k e^{K} = e^{2K}, \qquad (28)$$

while

$$K = \alpha \quad \text{and} \quad \left(\frac{\partial}{\partial T}\right)^i K = 0, \quad \text{when} \quad T = 0 \quad (1 \leqslant i \leqslant k-1). \qquad (29)$$

The two most important special cases are, of course, (i) $k = 1$, and (ii) $k \to \infty$. In case (i) let $e^{K} \equiv M$; then the differential equation becomes

$$\frac{\partial M}{\partial T} = M(M-1), \quad \text{with} \quad M(0) = e^{\alpha},$$

which is easily found to have the solution

$$M = \frac{1}{1-e^{T}(1-e^{-\alpha})},$$

in agreement with the results of Feller mentioned at the end of § 2. On the other hand, in case (ii), as $k \to \infty$ the operator

$$\left(1+\frac{1}{k}\frac{\partial}{\partial T}\right)^k$$

formally approaches the limit $\exp\left(\frac{\partial}{\partial T}\right)$,

and so the equation (28) takes the form

$$K(\alpha, T+1) = 2K(\alpha, T)$$

with the initial conditions

$$K(\alpha, 0) = \alpha \quad \text{and} \quad \left(\frac{\partial}{\partial T}\right)^i K(\alpha, 0) = 0 \quad (\text{all } i \geqslant 1).$$

The solution is $K(\alpha, T) = \alpha 2^{[T]}$,

in agreement with the results for the deterministic model.

5. **The variance of n.** From the equation satisfied by the cumulant-generating function one can readily obtain an equation to determine the variance of the population size n. Since

$$e^K = 1 + \alpha\kappa_1 + \tfrac{1}{2}\alpha^2(\kappa_2+\kappa_1^2) + \ldots,$$

where the cumulants κ_1, κ_2, etc., are functions of θ, it follows that

$$\left(1+\frac{d}{d\theta}\right)^k \kappa_1 = 2\kappa_1, \tag{30}$$

and

$$\left(1+\frac{d}{d\theta}\right)^k (\kappa_2+\kappa_1^2) = 2(\kappa_2+\kappa_1^2) + 2\kappa_1^2. \tag{31}$$

The first of these equations provides an alternative starting-point for the investigation of the mean value of n, given in § 3, while the second is the required formula determining the variance. It is to be combined with the boundary conditions

$$\left(\frac{d}{d\theta}\right)^i \kappa_2(0) = 0 \quad (0 \leqslant i \leqslant k-1).$$

In principle the equation (31) could be solved explicitly, but fortunately there is little to be gained in carrying out this laborious task. It is enough to notice, first, that when the time t is large the dominating term in the complementary function will be a constant multiple of

$$e^{(2^{1/k}-1)\theta}, \tag{32}$$

while the dominating term in the particular integral is

$$\frac{2}{\left(1+\frac{2}{k}\alpha_k\right)^k - 2}\left\{\frac{2^{1/k}}{2\alpha_k}e^{(2^{1/k}-1)\theta}\right\}^2. \tag{33}$$

Thus, since (32) is of a smaller order than (33) when θ is large, it follows that

$$\text{Variance}\,(n) \simeq \left\{\frac{2}{\left(1+\frac{2\alpha_k}{k}\right)^k - 2} - 1\right\}\bar{n}^2, \quad = C_k\bar{n}^2, \tag{34}$$

as the time t tends to infinity. The values of C_k for the first few values of k are given below:

k =	1	2	3	4	5
C_k =	1·000	0·489	0·324	0·242	0·193

If k is allowed to approach infinity,

$$C_k \simeq \frac{2(\log 2)^2}{k}, \quad = \frac{0{\cdot}9609}{k};$$

the table shows that this asymptotic formula is true with surprising accuracy even for quite small values of k (the error when $k = 1$ being only 4 %). Accordingly the relation

$$\text{C. of V.}(n) \simeq \sqrt{\frac{2}{k}} \log 2, \tag{35}$$

for the coefficient of variation of the population size can be used for all values of k, provided that $\bar{n}$ is sufficiently large. In the following section an extension of (35) will be found, which can be used whenever *one* of k and $\bar{n}$ is large enough. Some values of (35) are tabulated below (although the first entry has been based on the exact value $C_1 = 1$).

C. *of* V. (n), *for large* $\bar{n}$

k	$= 1$	5	20	100
C. of V. (n) =	100 %	44 %	22 %	10 %

6. **The asymptotic form for the distribution of the population size when the parameter k tends to infinity.** Once again it is convenient to write $\theta = kT$ and $e^K = M$, so that $T = \lambda t$; $M(\alpha, T)$ is then the moment-generating function for the population size n, and satisfies the differential equation

$$\left(1 + \frac{1}{k}\frac{\partial}{\partial T}\right)^k M = M^2, \tag{36}$$

with the boundary conditions

$$M = e^\alpha \quad \text{and} \quad \left(\frac{\partial}{\partial T}\right)^i M = 0 \quad (1 \leqslant i \leqslant k-1), \quad \text{when} \quad T = 0.$$

It was seen in the last section that for large T and k the coefficient of variation of n is of order $k^{-\frac{1}{2}}$; this suggests that it may be appropriate to consider the standardized variable

$$X \equiv k^{\frac{1}{2}}(n - 2^T), \tag{37}$$

which one might expect to have a non-trivial limiting distribution as k tends to infinity, for any fixed value of t. The variable X has the moment-generating function

$$E(e^{\beta X}) \equiv M_0(\beta, T) = \exp(-k^{\frac{1}{2}}\beta 2^T) M,$$

where M is the solution of (36) associated with the boundary conditions

$$M = \exp(\beta k^{\frac{1}{2}}) \quad \text{and} \quad \left(\frac{\partial}{\partial T}\right)^i M = 0 \quad (1 \leqslant i \leqslant k-1), \quad \text{when} \quad T = 0.$$

The problem is, therefore, to substitute

$$\exp(k^{\frac{1}{2}}\beta 2^T) M_0(\beta, T)$$

for M in (36), and then to find the asymptotic form (as k tends to infinity) of the solution satisfying the boundary conditions

$$M_0 = 1 \quad \text{and} \quad \left(\frac{\partial}{\partial T}\right)^i M_0 = \exp(\beta k^{\frac{1}{2}}) \left(\frac{\partial}{\partial T}\right)^i \exp(-k^{\frac{1}{2}}\beta 2^T) \quad (1 \leqslant i \leqslant k-1), \quad \text{when} \quad T = 0. \tag{38}$$

The formal solution to this problem will now be given. A rigorous treatment would be preferable, but one does not immediately suggest itself; it therefore seems worth while pursuing the investigation by heuristic methods.

In the first place, it is convenient to write

$$\Lambda(x) \equiv e^{-x}\left(1+\frac{x}{k}\right)^k \equiv \sum_{m=0}^{\infty} f_m(k)\frac{x^m}{k^{m-1}}, \tag{39}$$

and then to write the fundamental equation (36) in the form

$$\Lambda\left(\frac{\partial}{\partial T}\right) M(T) = \{M(T-1)\}^2,$$

or, on expansion,

$$\sum_{m=0}^{\infty} \frac{f_m(k)}{k^{m-1}}\left(\frac{\partial}{\partial T}\right)^m M(T) = \{M(T-1)\}^2. \tag{40}$$

The first few coefficients $f_m(k)$ are

$$f_0(k) = 1/k, \qquad f_1(k) = 0,$$
$$f_2(k) = -\tfrac{1}{2}, \qquad f_3(k) = +\tfrac{1}{3},$$
$$f_4(k) = -\tfrac{1}{4}+\tfrac{1}{8}k, \quad f_5(k) = +\tfrac{1}{5}-\tfrac{1}{6}k,$$

and on considering the mode of formation of the general term of the expansion* it will be found that when $m = 2p+1$ the largest term in $f_m(k)$ is of order k^{p-1}, while when $m = 2p$ the largest term is

$$\frac{(-)^p}{2^p p!}k^{p-1}. \tag{41}$$

On the other hand, for all values of m the largest term in

$$\left(\frac{\partial}{\partial T}\right)^m \{\exp(k^{\frac{1}{2}}\beta 2^T) M_0(\beta, T)\},$$

as k tends to infinity (supposing M_0 and its derivatives to remain of finite order in these circumstances), is

$$k^{\frac{1}{2}m}(\beta 2^T \log 2)^m \exp(k^{\frac{1}{2}}\beta 2^T) M_0(\beta, T).$$

Thus if

$$\lim_{k\to\infty} M_0(\beta, T) = M_1(\beta, T),$$

it will follow that

$$\frac{\{M_1(\beta, T-1)\}^2}{M_1(\beta, T)} = \sum_{p=0}^{\infty}\frac{(-)^p}{p!}\{\tfrac{1}{2}\beta^2(\log 2)^2 2^{2T}\}^p$$
$$= \exp\{-\tfrac{1}{2}\beta^2(\log 2)^2 2^{2T}\}, \tag{42}$$

and the most general solution of this functional equation (when T is an integer) is easily seen to be

$$M_1(\beta, T) = \exp\{\beta^2(\log 2)^2 2^{2T} + A\, 2^T\},$$

where the constant A is to be determined from the initial conditions. Now

$$M_1(\beta, 0) = M_0(\beta, 0) = 1,$$

and so when T is an integer,

$$M_1(\beta, T) = \exp\{\tfrac{1}{2}\beta^2 2(\log 2)^2 2^T(2^T-1)\}. \tag{43}$$

But this is the limiting form of the moment-generating function for the standardized random variable (37). Accordingly the above argument indicates that *as k tends to infinity the population size n is asymptotically normally distributed with the mean value*

$$\bar{n} = 2^{\lambda t} \tag{44}$$

and the variance

$$\frac{2(\log 2)^2}{k}\bar{n}(\bar{n}-1), \tag{45}$$

for every fixed integer value of λt.

* It is advisable to calculate the $f_m(k)$ from the successive expansion of $\exp\{\log\Lambda(x)\}$.

The rather unexpected agreement with formula (35) of § 5 will be noticed. It appears, in fact, that the approximate formula,

$$\text{C. of V.}(n) \simeq \log 2 \left(\frac{2}{k}\right)^{\frac{1}{2}} \left(1-\frac{1}{\bar{n}}\right)^{\frac{1}{2}}, \tag{46}$$

is valid whenever *one* of k and $\bar{n}$ is sufficiently large. The limiting values of (46), for large $\bar{n}$, are illustrated by the second table in § 5.

The purely formal character of the preceding argument is clearly responsible for the difference between (44) and the earlier, precise, result (23a). It is to be expected that the result (45) may be similarly incomplete, and a more careful analysis would be of interest.

7. **The coefficient of variation of the population size for more general types of stochastic birth process.** It has been shown in § 5 that for the multiple-phase birth process

$$\text{C. of V.}(n) \simeq A_k/k^{\frac{1}{2}}, \quad \text{when} \quad \bar{n} \text{ is large,}$$

where the coefficient A_k is equal to unity when $k = 1$ and has the limit 0·98 as k approaches infinity. On the other hand, it easily follows from the definition of the process that

$$\text{C. of V.}(\tau) = 1/k^{\frac{1}{2}}$$

(where τ, as usual, denotes the generation time). The similarity of the two results is striking, and their relationship will be made clearer by the following crude argument, a precise formulation of which is yet to be found. This gives reasons for supposing that the relation,

$$\text{C. of V.}(n) \simeq 0{\cdot}98\ \text{C. of V.}(\tau), \quad \text{when} \quad \bar{n} \text{ is large,} \tag{47}$$

is true for *all* stochastic birth processes of the general type discussed in § 1, *whatever the form of the distribution of generation time*, provided that C. of V. (τ) is small enough for the process to be equivalent, in regard to its mean growth, to the deterministic process for which

$$n = 2^{\lambda t},$$

where $1/\lambda$ is the mean generation time for the *actual* process. To cast the subsequent argument into a rigorous form it would be necessary to pay careful attention to the 'spreading' of generations; this will not be attempted here, and the conclusions should be regarded as tentative only.

Suppose then that t (and so also $\bar{n}$) is so large as to make it 'practically certain' that the first g generations have been completely established.*

Let $\tau_{ij} \quad (i = 1, 2, \ldots, g-1; j = 1, 2, \ldots, 2^{i-1})$

be the time which elapses between the 'birth' of an individual (to be identified by the suffix j) in the ith generation and its own later subdivision. Thus, for example, τ_{11} is the epoch at which the initial individual subdivides, to be replaced by the two individuals of the second generation which themselves subdivide at the epochs $\tau_{11}+\tau_{21}$ and $\tau_{11}+\tau_{22}$ respectively. When the gth generation has been established its 2^{g-1} members will continue to generate 2^{g-1} independent subpopulations during the intervals of time severally left to them until the final count is taken at the epoch t. If any one member of the gth generation is 'born' at the epoch u, it will have a time $(t-u)$ available for continued subdivision, and the size of the subpopulation which it generates will be (in the deterministic model)

$$2^{\lambda(t-u)}.$$

* One (artificial) way of making this statement precise would be to impose an upper bound to the possible values of the generation time. This, however, would exclude the multiple-phase process.

It is a principle of the present calculation that once the first-order fluctuations have been identified their coefficients can be calculated on the basis of the deterministic model. A fluctuation δu in u thus implies a fluctuation

$$-\lambda\,\delta u\,2^{\lambda(t-u)}\log 2$$

in the *expected* size of the subpopulation developed at the epoch t, and on inserting the 'deterministic' value for u in the coefficient, this becomes

$$-\lambda\,\delta u\,2^{\lambda t-g+1}\log 2.$$

Now δu is the sum of the fluctuations in the generation times for each of the specified individual's $(g-1)$ ancestors. If for the moment the process is imagined to develop strictly according to expectation once the gth generation has been established, it will be seen that the contribution of the fluctuation of τ_{ij} to the final population size n will be

$$-2^{g-1}\lambda\,\delta\tau_{ij}\,2^{\lambda t-g+1}\log 2,$$

multiplied by the fraction of individuals in the gth generation who possess (i,j) as an ancestor. This fraction is

$$1/2^{i-1},$$

and so

$$\frac{\partial n}{\partial \tau_{ij}} = -\lambda\,2^{\lambda t-i+1}\log 2. \tag{48}$$

The total contribution to $\mathrm{Var}\,(n)$ from the fluctuations possible during the development of the first g generations is thus

$$\lambda^2(\log 2)^2\,\mathrm{Var}\,(\tau)\,\{1(2^{\lambda t})^2+2(2^{\lambda t-1})^2+2^2(2^{\lambda t-2})^2+\ldots+2^{g-2}(2^{\lambda t-g+2})^2\}$$
$$=2\lambda^2(\log 2)^2\,\mathrm{Var}\,(\tau)\,2^{2\lambda t}\left(1-\frac{1}{2^{g-1}}\right).$$

To this must be added the further contribution arising from fluctuations in the development of the $(g+1)$th and later generations. Since this is equivalent to the growth of 2^{g-1} independent populations during a time $\left(t-\dfrac{g-1}{\lambda}\right)$, the additional contribution to the variance of n will be

$$2^{g-1}\,2\lambda^2(\log 2)^2\,\mathrm{Var}\,(\tau)\,2^{2\lambda t-2g+2},$$

to the first order (adopting the usual iterative procedure). If now $\bar{n}$ is large enough to justify the use of a sufficiently large g, it will follow that the second contribution is negligible and that

$$\text{C. of V.}\,(n)\simeq 2^{\frac{1}{2}}\log 2\ \text{C. of V.}\,(\tau),$$

as required. Since $2^{\frac{1}{2}}\log 2 = 0{\cdot}98$, it is therefore suggested as a practical rule-of-thumb that *the coefficients of variation for the population size and the generation time are approximately equal.* Of course, if the initial population size N is greater than unity this estimate of C. of V. (n) must be divided by $\sqrt{N}$, since each of the N initial members will generate independently a subpopulation to which the result will apply.

SUMMARY

This paper presents a mathematical account of a stochastic birth process in which the generation time (the interval between 'birth' and 'parenthood' in the life of an individual) is distributed like a χ^2-variate with $2k$ degrees of freedom. When $k=1$, the process reduces to the simple stochastic birth process in the form originally introduced by W. Feller (1939), while as k tends to infinity the process assumes a strictly deterministic form in which the

population undergoes an exact doubling at regular intervals. It is suggested that for intermediate values of k (say of the order of 20) this new 'multiple-phase' stochastic birth process represents a further step towards the construction of an adequate mathematical model of the growth of real populations of elementary organisms.

In order to show that the methods of the paper are relevant in an actual biological example, a brief discussion is given of the work of Kelly & Rahn (1932) on the distribution of generation times for *Bacterium aerogenes*.

The paper concludes with a sketch of a more general argument suggesting that for a wide class of stochastic birth processes the coefficient of variation of the population size is ultimately approximately equal to the coefficient of variation of the generation time, when this is sufficiently small for the process to approximate to the deterministic form, the population being assumed to have developed from a single 'ancestor' in the absence of 'mortality'.

In conclusion, I should like to express my thanks to Mr D. J. Finney for kindly making available to me the computing facilities at his disposal, during the preparation of this paper.

REFERENCES

ARLEY, N. & BORCHSENIUS, V. (1945). On the theory of infinite systems of differential equations and their application to the theory of stochastic processes and the perturbation theory of quantum mechanics. *Acta Math.* **76**, 261–322, esp. pp. 298–9.

BARTLETT, M. S. (1947). *Stochastic Processes.* (Notes of a course given at the University of North Carolina in the Fall Quarter, 1946.) It is understood that copies of these notes are available on request.

BARTLETT, M. S. & KENDALL, D. G. (1946). The statistical analysis of variance-heterogeneity, and the logarithmic transformation. *J. R. Statist. Soc. Suppl.*, **8**, 128–38.

FELLER, W. (1939). Die Grundlagen der Volterraschen Theorie des Kampfes ums Dasein in wahrscheinlichkeitstheoretischer Behandlung. *Acta Biotheoretica*, **5**, 11–40.

KELLY, C. D. & RAHN, O. (1932). The growth rate of individual bacterial cells. *J. Bacteriol.* **23**, 147–53.

KENDALL, D. G. (1947). A review of some recent work on discontinuous Markoff processes with applications to biology, physics, and actuarial science. *J. R. Statist. Soc.* **110**, 130–7.

KENDALL, D. G. (1948*a*). On some modes of population growth leading to R. A. Fisher's logarithmic series distribution. *Biometrika*, **35**, 6–15.

KENDALL, D. G. (1948*b*). On the generalized 'birth-and-death' process. *Ann. Math. Statist.* **19**, 1–15.

2×2 TABLES; THE POWER FUNCTION OF THE TEST ON A RANDOMIZED EXPERIMENT

BY E. S. PEARSON AND MAXINE MERRINGTON

1. INTRODUCTORY

In his discussion of significance tests for 2 × 2 tables, Barnard (1947) has pointed out how data classified in the form of Table 1 may appear as the outcome of a number of different types of investigation. Differences in point of view which have been expressed regarding the handling of the figures, concern the probability constructs by aid of which the bare numerical data recorded in the table provide a basis for inference. Two lines of approach may be distinguished.

Table 1

	Col. 1	Col. 2	Total
Row 1	a	c	m
Row 2	b	d	n
Total	r	s	N

Following the first, it is considered that for all the types of problem,* the relevant information on the points at issue may be obtained by comparing the observed pattern of cell contents (a, b, c, d) with the set of possible patterns, all giving the marginal totals actually found in the sample. Thus there is only one degree of freedom among the four cell frequencies, and the relevant probability distribution is obtained from the hypergeometric series. This approach can be derived from Fisher's information theory. But without using this theory, a may be referred to the one-dimensioned, conditional set as a convenient practical device, which avoids the introduction of nuisance parameters.

From the point of view of the second approach it is an over-simplification to treat every case providing data in the form of Table 1 as a problem of sampling with fixed marginal totals. It is suggested that the readiness of the mind to assimilate the information provided by the statistical analysis depends on the directness of the relation between the theoretical probability set and the random process of selection introduced in collecting the data. Since the random procedure may have entered in different ways, the appropriate probability constructs may be expected to differ.

It can be argued that, except in the case of small samples, the difference of approach is practically unimportant, and that even here, until tables of the kind which Barnard has in mind are available, the statistician will be forced to draw his conclusions from a table of the conditional distribution, such as that recently prepared by Finney (1948). But there is another aspect to the matter. The published discussion has hitherto been concerned primarily with the sampling distribution of a statistic under the null hypothesis. If we go beyond this

* Three types were discussed by Barnard (1947) and Pearson (1947).

and consider the sensitivity of the test, that is to say, its power to detect differences if they exist, it is at once clear that all problems cannot be treated in the same manner.

In a recent paper, Patnaik (1948) has considered this aspect of what one of us (Pearson, 1947) has termed Problem II and of what Barnard termed the 2×2 comparative trial. This occurs when we inquire whether the probability of an individual bearing a given character A, is the same in two large populations from which random samples of size m and n, respectively, have been drawn. Here, two separate random selections are involved, and the cell contents may be described as having two degrees of freedom. If the two population probabilities are unequal, i.e. $p_1(A) \neq p_2(A)$, the chance that the test will establish a difference at a given significance level α is a function of p_1 and p_2, of m and n and of α. This relationship was explored by Patnaik.

In the present paper we shall consider what Pearson termed Problem I and Barnard the 2×2 independence trial, from this aspect of the power of the test. In this case, only a single process of random selection or partition is called for.

2. Statement of the Problem

For convenience we shall describe the type of experiment we have in mind as one in which two 'treatments', say A and B, are compared; the response is a quantal one, so that an individual either 'reacts' or 'fails to react'. The applications suggested by these terms lie in the biological field, but there is no difficulty in translating the terms of the theoretical picture to fit a case where, for example, the individual is a shell, the two treatments are two types of fuze and the reaction is successful perforation of a steel plate.

In this Problem I, the N individuals available for experiment are divided by a random partition into a group of m which receive treatment A, and a group of $n = N - m$ which receive treatment B. It is then observed that a/m and b/n react in the specified manner. The experiment is self-contained and the random process under complete control; but without further assumptions or knowledge, the inferences that are possible relate only to the reactions of the N individuals to the treatments. This may be all that is called for. Inferences of wider application may be drawn by assuming that the N have been sampled randomly from a population in which we are interested. Or, as is often the case, the experiment may be one of a related series, each experiment in which is self-contained. These, taken as a whole, can form the basis of reasoned conclusions regarding the treatments, conclusions which are not dependent on all groups of individuals having been drawn randomly from a unique population in the rigorous statistical sense. This is the case if the tests are applied to laboratory animals whose susceptibility may change somewhat from time to time. Or, as Barnard has suggested, when an open-air gunnery trial runs over several days of inconstant weather.

The question then is this: confining attention to the group of N individuals, in what sense is it possible to interpret a difference in treatments and how can we measure the power of the test to detect such a difference? Perhaps the most general method of regarding the problem is to suppose that the N individuals fall into four classes:

(i) those who would react if given either treatment, X in number;
(ii) those who would react only if given treatment A, W in number;
(iii) those who would react only if given treatment B, Y in number;
(iv) those who would react to neither treatment, Z in number.

In this way we recognize that every individual does not respond in the same way to a given treatment; that the success of a shell in perforating a plate will depend not only on the fuzing, but on other factors such as strength of shell-case, angle of yaw on striking, the position of the strike on the plate, etc. These latter factors are purposely randomly associated with the two types of fuze under trial.

If m individuals are selected randomly and assigned treatment A and the remaining n assigned treatment B, the resulting partition will be that shown in Table 2a; as, however, we can only observe whether an individual has reacted or not, the figures available for analysis will be in the form of Table 2b.

Table 2a

	React if given A or B	Only react if given A	Only react if given B	React to neither treatment	Total
Treatment A	x_1	w_1	y_1	z_1	m
Treatment B	x_2	w_2	y_2	z_2	n
Total	X	W	Y	Z	N

Table 2b

	React	Fail to react	Total
Treatment A	$x_1+w_1=a$	$y_1+z_1=c$	m
Treatment B	$x_2+y_2=b$	$w_2+z_2=d$	n
Total	$X+w_1+y_2=r$	$Z+w_2+y_1=s$	N

The usual null hypothesis is that the treatments are identical as far as producing a reaction on these N individuals is concerned, i.e. it is the hypothesis that $W = Y = 0$. The test of significance of departure from hypothesis would be applied to the 2×2 Table 2b, and in its exact form consists in referring $a = x_1 + w_1$ to the appropriate hypergeometric distribution, with parameters r, N and m. Here, if the sample is not too large, Finney's (1948) table is applicable. More approximately, we can either regard

$$\frac{a - rm/N}{\sqrt{\left(\frac{mnrs}{N^2(N-1)}\right)}}$$ as a unit normal deviate, or

$$\frac{(ad-bc)^2 N}{mnrs}$$ as a χ^2 with 1 degree of freedom,

making a correction for continuity, if necessary.

It should be noted that if treatment A were regarded as more successful than B when $X + W > X + Y$, then the null hypothesis to test would be that $W - Y = 0$. The expectation of the difference $a/m - b/n$ is still zero if $W = Y \neq 0$, but its sampling distribution under random partition can no longer be determined from the marginal totals of Table 2*b*. The experiment as planned is not, in fact, able to distinguish between the cases $W = Y = 0$ and $W - Y = 0$; to do so would involve applying both treatments to the same individuals, which will usually be impossible in practice.

Table 3*a*

	React if given A or B	React if given B	React to neither treatment	Total
Treatment A	x_1	y_1	z_1	m
Treatment B	x_2	y_2	z_2	n
Total	X	Y	Z	N

Table 3*b*

	React	Fail to react	Total
Treatment A	$x_1 = a$	$y_1 + z_1 = c$	m
Treatment B	$x_2 + y_2 = b$	$z_2 = d$	n
Total	$X + y_2 = r$	$Z + y_1 = s$	N

In the discussion which follows we shall suppose that $W = 0$. This may narrow the field of application, but not too seriously. If A is an old treatment and B a new one which it is hoped is an improvement,* we are assuming that B has at least the qualities of A. Thus if B aims at the cure of a disease and A is a control corresponding to 'no treatment', we assume that in no case will B *prevent* a recovery which would have taken place without any treatment at all. With $W = 0$, we have the scheme of Tables 3*a* and *b*. In this case the null hypothesis is that $Y = 0$; the alternative that interests us is that $Y > 0$, so that the statistical test applied to Table 3*b* involves comparing the observed value of a with its lower significance level, or b with its upper level. This critical limit for a may be written $a(\alpha, r, N, m)$, where α is the chance of falling at or below the limit if $Y = 0$. It can be determined either precisely from the hypergeometric or from the normal approximation. If the null hypothesis is not true, $r = X + y_2$ will vary from one random partition to another. Thus the problem is to determine the probability that, when $Y \neq 0$, $a \leqslant a(\alpha, r, N, m)$, where both a and r are random variables.

* For purposes of discussion, we take 'reaction' to be good.

A numerical illustration will help to make the position clear. Thirty small disks were placed in a box, of which $X = 10$ were coloured red, $Y = 10$ coloured green and $Z = 10$ coloured yellow. The disks were divided randomly into two groups A and B each containing $m = n = 15$. In terms of treatment comparisons, reds 'respond to treatment' if they fall into group A, reds and greens if they fall into group B. Four of many possible partitions are shown in the 2×3 tables of row I below; under these are given in row II the corresponding 2×2 tables which contain the numerical data available to the experimenter. He cannot distinguish between greens and yellows in group A or between reds and greens in group B. The null hypothesis is that there were no green disks in the box of 30. Below the 2×2 tables are shown (III) the critical limits corresponding to a nominal 5 % significance level, (IV) the true level or sum of the tail terms and (V) the conclusion that would be drawn. It may be noted that this experimental partition was repeated 50 times and that significance was established, i.e. the presence of green disks inferred, in 20 of these partitions. This is a result which, as will be shown below, agrees closely with expectation.

	Colour	...	R.	G.	Y.		R.	G.	Y.		R.	G.	Y.		R.	G.	Y.	
I	A		3	6	6	15	5	5	5	15	5	4	6	15	7	3	5	15
	B		7	4	4	15	5	5	5	15	5	6	4	15	3	7	5	15
			10	10	10	30	10	10	10	30	10	10	10	30	10	10	10	30
II				3	12	15		5	10	15		5	10	15		7	8	15
				11	4	15		10	5	15		11	4	15		10	5	15
				14	16	30		15	15	30		16	14	30		17	13	30
III	Rejection level, given r		$a_0 = 4$				$a_0 = 4$				$a_0 = 5$				$a_0 = 5$			
IV	$P\{a \leqslant a_0 \mid r\}$		0·0328				0·0134				0·0328				0·0127			
V	Significant		Yes				No				Yes				No			

The probability that $a \leqslant a(\alpha, r, N, m)$ when $Y \neq 0$ represents the power of the test in the sense of Neyman & Pearson, i.e. the probability of establishing significance at the 100α % level when $Y > 0$. It will be a function not only of Y but of X (or Z). Owing to discontinuity in the distribution of a, it is, of course, impossible to choose $a(\alpha, r, N, m)$ with the same value of α for all r. In our calculations given below we have used what may be termed nominal 5 and 15 % levels, such that the critical limits $a(\alpha, r, N, m)$ are the highest integer values satisfying the inequalities

$$P\{a \leqslant a(\alpha, r, N, m)\} \leqslant \alpha \quad (\alpha = 0{\cdot}05 \text{ and } 0{\cdot}15). \tag{1}$$

We have taken $m = n = \frac{1}{2}N$, supposing that each treatment is given to the same number of individuals. This will usually be the case in a planned experiment with randomization, and as our object is to illustrate certain points which we believe are of general interest, no exhaustive tabulation is called for.

The method of calculation leading to Figs. 2–7 is described in the following section and a discussion of results is given in § 4.

3. Method of Calculation

In a 2×3 table of fixed margins, there are two degrees of freedom. We shall take as independent variables x_1 and y_1. The probability of a particular partition of Table 3*a* is

$$p(x_1, y_1) = \frac{m!\,n!\,X!\,Y!\,Z!}{x_1!\,x_2!\,y_1!\,y_2!\,z_1!\,z_2!\,N!}, \tag{2}$$

where x_2, y_2, z_1 and z_2 can all be expressed in terms of x_1, y_1 and the marginal totals. For mathematical convenience we may regard the partition of Table 3*a* as obtained in two steps, the first determining y_1, the second x_1. In fact, the expression (2) may be written as a product of two parts

$$\frac{m!\,n!\,Y!(N-Y)!}{y_1!\,y_2!(m-y_1)!\,(n-y_2)!\,N!} \times \frac{X!\,Z!(x_1+z_1)!\,(x_2+z_2)!}{x_1!\,x_2!\,z_1!\,z_2!(X+Z)!}. \tag{3}$$

The first factor, say $p(y_1)$, is the probability of obtaining y_1 individuals with a character A in a sample of Y drawn randomly, without replacement, from a population of N individuals of whom m have a character A. The second factor, say $p(x_1 \mid y_1)$, is the probability of obtaining x_1 with A in a sample of X drawn, without replacement, from the remaining $N-Y$ individuals, of whom $m-y_1$ now have A. Thus

$$p(x_1 y_1) = p(y_1) \times p(x_1 \mid y_1), \tag{4}$$

and while $p(y_1)$ is the marginal distribution for y_1 of the joint distribution $p(x_1 y_1)$, $p(x_1 \mid y_1)$ is the relative probability distribution of x_1 in the array of constant y_1. Both $p(y_1)$ and $p(x_1 \mid y_1)$ are of hypergeometric type; the joint distribution has been discussed by K. Pearson (1924) who termed it a double hypergeometric series and considered how it might lead to a frequency surface.* He gave some of the momental constants of $p(x_1 \mid y_1)$ which we shall use below; in our present notation, and for the special case $m = n = \frac{1}{2}N$, these become

$$\text{Mean } (x_1 \mid y_1) = Xp, \tag{5}$$

$$\text{Variance } (x_1 \mid y_1) = Xpq\frac{Z}{X+Z-1}, \tag{6}$$

$$\beta_1(x_1 \mid y_1) = \frac{(p-q)^2\{1-2(X-1)/(X+Z-2)\}^2}{Xpq \quad 1-(X-1)/(X+Z-1)}, \tag{7}$$

where

$$p = 1-q = (m-y_1)/(N-Y) = (x_1+z_1)/(X+Z). \tag{8}$$

The distribution $p(x_1 y_1)$ is bounded by certain limiting lines, thus:

$$x_1 \leqslant X, \quad y_1 \leqslant Y, \quad m \geqslant x_1+y_1 \geqslant m-Z. \tag{9}$$

Fig. 1 illustrates the position for the case

$$X = 10, \quad Y = 15, \quad Z = 5, \quad m = n = \tfrac{1}{2}N = 15. \tag{10}$$

In applying the test of significance to the 2×2 Table 3*b*, we shall reject the null hypothesis ($Y = 0$) when

$$x_1 = a \leqslant a(\alpha, r, N, m). \tag{11}$$

But

$$r = X+y_2 = X+Y-y_1. \tag{12}$$

Hence for given X and Y there will be a critical limit in each y_1 array, such that if x_1 fall below the limit, the hypothesis will be rejected. For the marginal values given in (10),

* He had illustrated the distribution many years before (K. Pearson, 1895) in connexion with the correlation between the number of cards of a given suit held by two players at whist.

Fig. 1 shows these rejection limits as a connected line for the nominal level $\alpha = 0{\cdot}05$.* The power of the test for a given X and Y is then the sum of the probabilities $p(x_1 y_1)$ taken over the points of the field outside (or below) this critical boundary. Fig. 4 shows that for this particular case the sum is about 0·85.

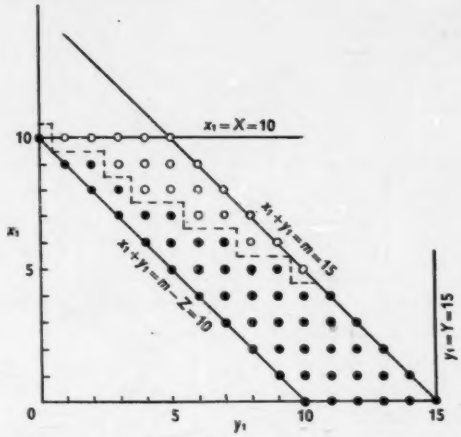

Fig. 1. Illustration for case of partition of 30 individuals into two groups, $m = n = \frac{1}{2}N = 15$; $X = 10$, $Y = 15$, $Z = 5$; nominal significance level, $\alpha = 0{\cdot}05$. The null hypothesis will be rejected at the (x_1, y_1) points marked ●, where $x_1 = a$, $y_1 = 25 - r$.

For small samples the calculations are not very tedious if undertaken systematically, but they become so when N is large. In the case $N = 50$ we have therefore used an approximation based on assuming that the array distributions $p(x_1 \mid y_1)$ can be represented by normal curves with the means and variances of equations (5) and (6). Since, using (8), when $m = \frac{1}{2}N$,

$$q - p = (2y_1 - Y)/(N - Y),$$

where the expectation of y_1 is $\frac{1}{2}Y$, it follows that the β_1 of the array distributions which contain most of the frequency will be small provided that neither X or Z are too small. The accuracy of this normal approximation is illustrated below.

The calculations on which Figs. 2–7 have been based may be summarized as follows:

(i) $N = 20$ *and* 30

Full calculations of the terms of the hypergeometric series

$$\frac{(m!)^2 r!(N-r)!}{a!b!c!d!N!}$$

were made for all values of r, and hence the levels corresponding to the nominal $\alpha = 0{\cdot}05$ and 0·15 were found. This determined the position of the boundary of the critical region as illustrated in Fig. 1 and also provided the terms of the series $p(y_1)$. The expressions $p(x_1 y_1)$ corresponding to points in the critical region were then found and summed for selected values of X, Y and Z, giving the ordinates of the power curves drawn in Figs. 2–5.

* These limits will be found to agree with those in Finney's (1948) table. The relevant section is that for $A = 15 = B$ (in his notation); further, since he takes $a > b$ his (i) a and (ii) b correspond to our (i) $25 - (x_1 + y_1)$ and (ii) x_1. See Appendix, p. 345 below. For example, points in Fig. 1 on the diagonal $x_1 + y_1 = 12$ make Finney's $a = 13$ and for this value he gives the 5 % significance level as $b = 7$; thus all points on this diagonal for which $x_1 \leqslant 7$ fall in the rejection region.

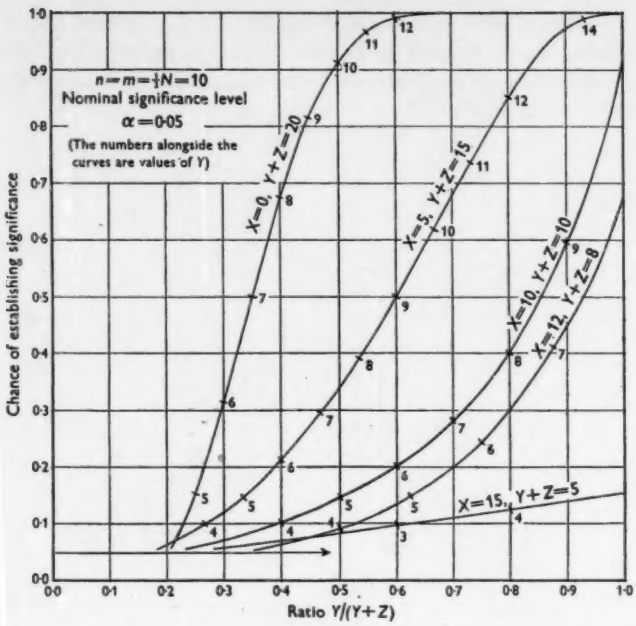
n=m=½N=10
Nominal significance level
α=0·05
(The numbers alongside the curves are values of Y)
X=0, Y+Z=20
X=5, Y+Z=15
X=10, Y+Z=10
X=12, Y+Z=8
X=15, Y+Z=5
Chance of establishing significance
Ratio Y/(Y+Z)

Fig. 2

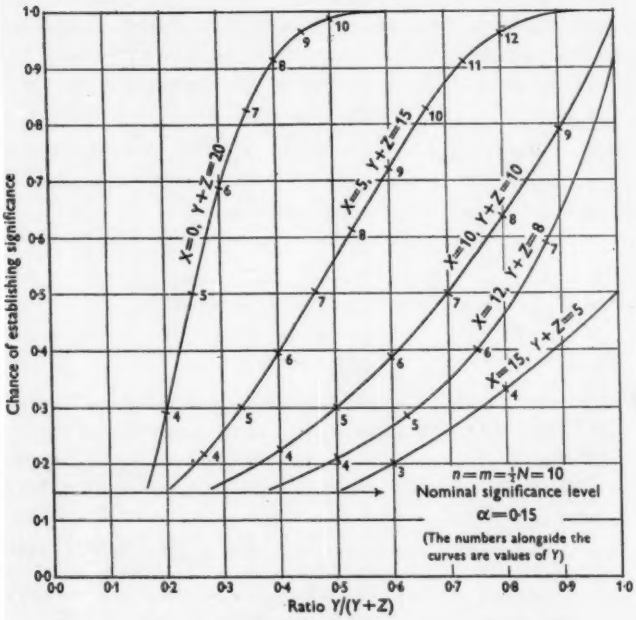
X=0, Y+Z=20
X=5, Y+Z=15
X=10, Y+Z=10
X=12, Y+Z=8
X=15, Y+Z=5
n=m=½N=10
Nominal significance level
α=0·15
(The numbers alongside the curves are values of Y)
Chance of establishing significance
Ratio Y/(Y+Z)

Fig. 3

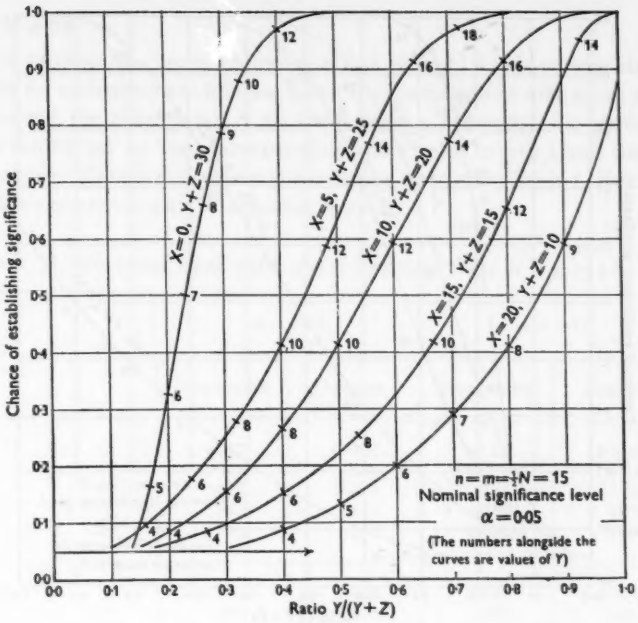
Chance of establishing significance
Ratio Y/(Y+Z)
X=0, Y+Z=30
X=5, Y+Z=25
X=10, Y+Z=20
X=15, Y+Z=15
X=20, Y+Z=10
n=m=½N=15
Nominal significance level
α=0·05
(The numbers alongside the curves are values of Y)

Fig. 4

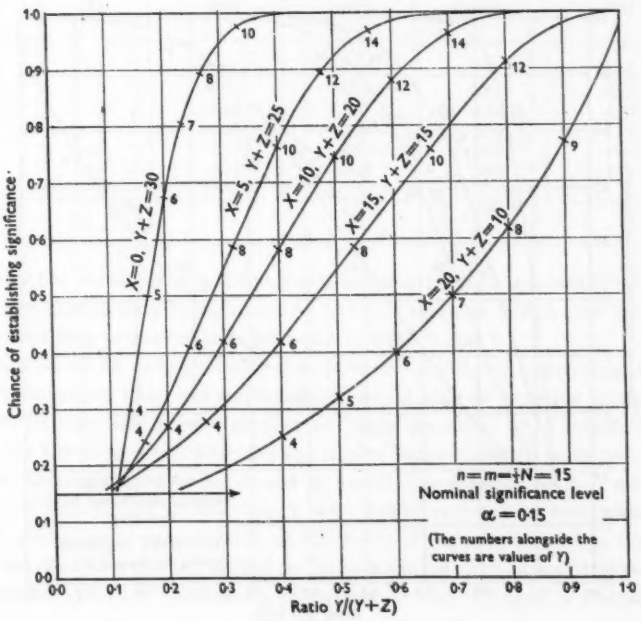
Chance of establishing significance
Ratio Y/(Y+Z)
X=0, Y+Z=30
X=5, Y+Z=25
X=10, Y+Z=20
X=15, Y+Z=15
X=20, Y+Z=10
n=m=½N=15
Nominal significance level
α=0·15
(The numbers alongside the curves are values of Y)

Fig. 5

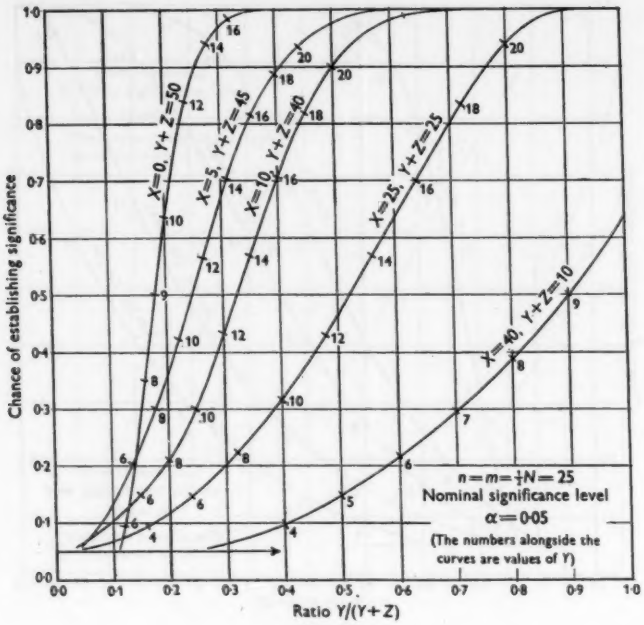
Chance of establishing significance
Ratio Y/(Y+Z)
X=0, Y+Z=50
X=5, Y+Z=45
X=10, Y+Z=40
X=25, Y+Z=25
X=40, Y+Z=10
n=m=½N=25
Nominal significance level
α=0·05
(The numbers alongside the curves are values of Y)

Fig. 6

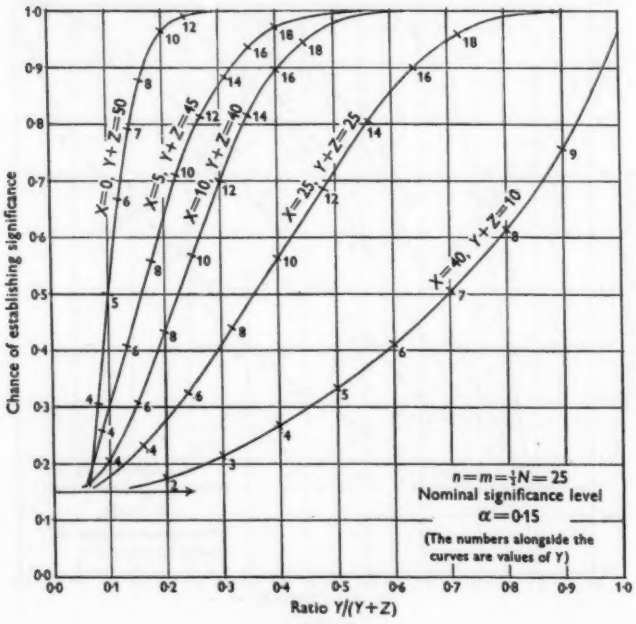
Chance of establishing significance
Ratio Y/(Y+Z)
X=0, Y+Z=50
X=5, Y+Z=45
X=10, Y+Z=40
X=25, Y+Z=25
X=40, Y+Z=10
n=m=½N=25
Nominal significance level
α=0·15
(The numbers alongside the curves are values of Y)

Fig. 7

(ii) *Check for* $N = 30$

For a number of cases the normal approximation referred to above was also used. This involved finding for each array the integral under the normal curve with mean and variance, (5) and (6), beyond the point $a(\alpha, r, N, m)+0{\cdot}5$, using a correction for continuity. These integrals were multiplied by the corresponding exact value of $p(y_1)$ and summed for all arrays. A comparison of exact and approximate results is shown in Table 4. From our present point of view the agreement may be regarded as good.

Table 4. *Approximate solution for power function. Case* $m = n = \frac{1}{2}N = 15$

X	Y	$\alpha = 0{\cdot}05$		$\alpha = 0{\cdot}15$	
		True power	Approx.	True power	Approx.
5	2	0·047	0·053	0·121	0·126
	5	·135	·139	·321	·324
	10	·413	·414	·764	·758
	15	·852	·844	·984	·979
	20	·999	·997	1·000	·999
10	2	0·039	0·043	0·166	0·170
	6	·155	·161	·420	·421
	10	·413	·415	·745	·740
	14	·773	·767	·962	·956
	18	·989	·989	1·000	1·000
15	1	0·033	0·036	0·136	0·140
	3	·064	·067	·224	·227
	6	·152	·157	·419	·420
	9	·326	·329	·676	·673
	12	·655	·651	·913	·908
	14	·951	·948	·990	·989

(iii) $N = 50$

Encouraged by the check for $N = 30$, the approximate method was employed throughout in this case, a few checks only being made by the full method, which now becomes rather laborious. The resulting power curves are shown in Figs. 6 and 7.

As N increases above 50, it is probable that further simplifying approximations could be introduced, but these we have not explored. It would also be possible to form diagrams giving contours of constant power, similar to those provided by Patnaik (1948) in his Figs. 2 and 3, for the case where samples are drawn independently from two populations. The parameters corresponding to his p_1 and p_2 would now seem to be X/N and $(X+Y)/N$, the proportion of individuals among the N who would respond to treatments A and B, respectively. A preliminary examination on the basis of the results used to form Figs. 2–7, suggests that the contours corresponding to various values of N, α and the power, may all belong approximately to a single family of curves, as was the case in Patnaik's problem, for $m = n$.

4. Discussion

In the first place some comment is necessary on the significance levels chosen. Owing to the discontinuity which occurs because the cell contents can assume integer values only, the standard levels of 10, 5, 1 %, etc., have not their ordinary meaning. As an example, take the case where $m = n = \frac{1}{2}N = 25$, $r = 16$. On the null hypothesis we have:

Chance that $a \leqslant 6$ is 0·182,
$\leqslant 5$ is 0·064,
$\leqslant 4$ is 0·016,
$\leqslant 3$ is 0·003.

Thus if we wish to accept *no more than* a 1 in 20 risk of rejecting the hypothesis when it is true, we should only do so when $a \leqslant 4$. But in practice, if prepared to accept a risk of *about* 1 in 20, we should clearly take $a = 5$ as the limit. Similarly, we should take $a = 4$ for a risk of *about* 1 in 100. Thus we should tend to base our conclusions on the actual tail sum found after the data are collected. But if we use the power function concept to inquire *in advance* how large N must be to make an experiment worth while, then we must think in terms of a specific upper limit or nominal level α. Hence it is important to know how much on the average the true level falls short of the nominal. Table 5 has been prepared to indicate this difference in the cases with $N = 20$, 30 and 50 and for nominal levels of $\alpha = 0{\cdot}05$ and 0·15. Finney's (1948) table also brings this out.*

These upper limit values are well on the safe side, and this may be what is wanted if we attach prime importance to Neyman & Pearson's first kind of error; i.e. to the risk of assuming a difference when none exists. But where it is of first importance to find a new solution, e.g. an improvement in treatment, higher risks in this direction will be accepted in order to avoid the chance of falling into the second kind of error, i.e. of overlooking a real difference when it exists. For this reason our calculations have been made for $\alpha = 0{\cdot}05$ and also for the rather high value $\alpha = 0{\cdot}15$, which, as can be found from Table 5, means on the average† a true level of $\bar{\alpha} = 0{\cdot}057$ for $N = 20$, $\bar{\alpha} = 0{\cdot}082$ for $N = 30$, $\bar{\alpha} = 0{\cdot}089$ for $N = 50$.

Turning to the interpretation of the diagrams, it must be emphasized again that the frequencies X, Y and Z will of course not be known; the purpose of the charts is to show for certain sample sizes, the combinations of these three frequencies needed to give a reasonable chance of establishing a significant treatment difference. If the investigation has the positive objective of establishing the value of new methods, it will naturally be hoped that the comparative experiment will establish statistical significance. If, with the knowledge available, it can be shown that an experiment of given magnitude is very unlikely to lead to a significant result, then it may be a waste of time to proceed on this scale. We think that the results presented in the diagrams will be helpful in this type of review.

If treatment A has already been in use, some information will be available as to the likely value of X/N, the proportion of individuals in a group of N who would react to A. We shall at least know whether it is more likely to be 0 or $\frac{1}{2}$. The ratio $Y/(Y+Z)$ used as the abscissa for the power curves measures the headway which the new treatment B could make in causing a satisfactory response among individuals with whom treatment A fails. Again, the experimenter will generally have some idea of what he hopes the treatment will achieve.

* See Appendix, p. 345 below.

† Giving equal weight to all values of r.

He may know, for example, that $Y/(Y+Z)$ is unlikely to exceed 0·25, and yet be clear that even if it were no larger than this the introduction of B would be amply justified. Here Figs. 4 and 5 show that, for $m = n = 15$, were $X = 0$ (treatment A ineffective*) and $Y/(Y+Z) = 8/30 = 0{\cdot}27$, the chance of establishing significance is about 0·66 at the nominal 5 % level and 0·89 at the nominal 0·15 % level. But if X were 10, or one-third of the group of $N = 30$ were responsive to the old treatment then, though 5 of the remaining 20 would

Table 5. *Showing for the case* $m = n = \frac{1}{2}N$: (i) *the critical limits* $a_0(\alpha, r, N, m)$ *for nominal significance levels* $\alpha = 0{\cdot}05$ *and* 0·15; (ii) *the true chance that* $a \leqslant a_0(\alpha, r, N, m)$

r	$N = 20$				$N = 30$				$N = 50$			
	$\alpha = 0{\cdot}05$		$\alpha = 0{\cdot}15$		$\alpha = 0{\cdot}05$		$\alpha = 0{\cdot}15$		$\alpha = 0{\cdot}05$		$\alpha = 0{\cdot}15$	
	a_0	True chance	a_0	True chance	a_0	True chance	a_0	True chance	a_0	True chance	a_0	True chance
3	—	—	0	0·105	—	—	0	0·112	—	—	0	0·117
4	0	0·043	0	0·043	0	0·050	0	0·050	—	—	0	0·055
5	0	0·016	0	0·016	0	0·021	0	0·021	0	0·025	0	0·025
6	0	0·005	1	0·070	0	0·008	1	0·084	0	0·011	1	0·095
7	1	0·029	1	0·029	1	0·040	1	0·040	1	0·049	1	0·049
8	1	0·010	2	0·085	1	0·018	2	0·107	1	0·024	2	0·123
9	2	0·035	2	0·035	1	0·007	2	0·054	1	0·012	2	0·069
10	2	0·012	3	0·089	2	0·025	3	0·123	2	0·037	3	0·144
11	—	—	—	—	2	0·010	3	0·064	2	0·019	3	0·085
12	—	—	—	—	3	0·030	4	0·132	3	0·048	3	0·048
13	—	—	—	—	3	0·013	4	0·070	3	0·025	4	0·098
14	—	—	—	—	4	0·033	5	0·136	3	0·013	4	0·057
15	—	—	—	—	4	0·013	5	0·071	4	0·031	5	0·108
16	—	—	—	—	—	—	—	—	4	0·016	5	0·064
17	—	—	—	—	—	—	—	—	5	0·038	6	0·119
18	—	—	—	—	—	—	—	—	5	0·021	6	0·072
19	—	—	—	—	—	—	—	—	6	0·042	7	0·124
20	—	—	—	—	—	—	—	—	6	0·023	7	0·077
21	—	—	—	—	—	—	—	—	7	0·044	8	0·128
22	—	—	—	—	—	—	—	—	7	0·024	8	0·079
23	—	—	—	—	—	—	—	—	8	0·046	9	0·131
24	—	—	—	—	—	—	—	—	8	0·025	9	0·081
25	—	—	—	—	—	—	—	—	9	0·046	10	0·131

For $r > \frac{1}{2}N$ the true chance for r is the same as that for $N-r$, while $a_0(\alpha, r, N, m) = r - m + a_0(\alpha, N-r, N, m)$.

react if given treatment B ($Y/(Y+Z) = 0{\cdot}25$), the experiment will most probably be inconclusive. This is because the power of the test is now only 0·12 and 0·34, respectively, at the 5 and 15 % levels.

If we regard $Y/(Y+Z)$ as the measure of effectiveness, the existence of X reduces the chance of establishing significance. We need not express effectiveness in this way, but the numerical data given in the charts associating power with the partition of N in X, Y and Z, makes it possible to express the position in any terms considered more appropriate.

As a further example, we may take the case where out of $N = 30$ individuals, $X = 10$ would react to treatment A and $X+Y = 20$ to treatment B. This might well be regarded

* At least, ineffective on the 30 individuals selected for the experiment.

as an eminently satisfactory result. But Figs. 4 and 5 show that the chance of distinguishing this case from one in which $Y = 0$ is only 0·41 at the 5 % level and 0·74 at the 15 % level.* If now we make a rough interpolation in Figs. 6 and 7 between the power curves for $X = 10$ and $X = 25$, it is seen that using 50 individuals for the experiment, then were $X = Y = 17$ (keeping the same proportions as before) there would be a chance of about 0·76 of establishing significance at the 5 % level and of about 0·92 at the 15 % level. There are likely to be circumstances where, with this knowledge, it would rightly be concluded that while a comparative experiment on two sets of 15 individuals would not be worth undertaking, an experiment using two sets of 25 would be.

Another point which the diagrams bring out is the dilemma which faces the medical research worker who wishes to establish his conclusions on a scientific basis, but has a strong conviction that a new treatment will be effective in reducing pain, hastening recovery or even saving life. Dealing with a hospital population of varying susceptibilities and with other changes in external conditions from time to time, it may be impossible to make comparisons which could be accepted indisputably between (*a*) successes in the past, using treatment A, and (*b*) successes at the moment using treatment B. Yet if a controlled, comparative test is carried out, it is seen that to provide a conclusive result the number of patients in the group Y must be considerable, perhaps 10, 16 or even 20. And yet, on the average, $\frac{1}{2}Y$ of these patients will be assigned the old treatment. Even if the belief in treatment B is based on intuition rather than evidence, considerations of ethics are likely to outweigh the urge to plan an experiment which makes a valid scientific comparison possible.

In conclusion, we freely admit that the presentation given in Figs. 2–7 cannot be regarded as a final one. It does, however, provide enough material for the statistician to consider whether the approach is practically useful and, hence, whether its extension and possible simplification are desirable.

REFERENCES

Barnard, G. A. (1947). *Biometrika*, **34**, 123.
Finney, D. J. (1948). *Biometrika*, **35**, 145.
Patnaik, P. B. (1948). *Biometrika*, **35**, 157.
Pearson, E. S. (1947). *Biometrika*, **34**, 139.
Pearson, K. (1895). *Philos. Trans.* A, **186**, 411.
Pearson, K. (1924). *Biometrika*, **16**, 172.

* In the experiment with disks referred to on p. 335 above, in 20 partitions out of 50 significance was established at the 5 % level and in 40 out of 50 at the 15 % level, results clearly consistent with these theoretical chances.

APPENDIX

Note on the arrangement of D. J. Finney's table (*Biometrika*, **35**, 145–56)

In the present paper we have been mainly concerned with a lower significance level for a, for a given value of $r = a+b$. This limiting value of a is determined by the sum of the tail terms in a hypergeometric series. Thus in Fig. 8 below, for $m = n = \frac{1}{2}N = 15$, when $r = 13$, what we have termed the nominal lower 5 % significance level for a is 3 and when $r = 12$ it is also 3. This is because the sum of the chances of a assuming values of 0, 1, 2 and 3 on the null hypothesis is 0·0127 in the former case and 0·0301 in the latter, while it would be over 0·05 if we included the chance that $a = 4$. On this basis, the stepped line indicates the position of the significance level for different values of r.

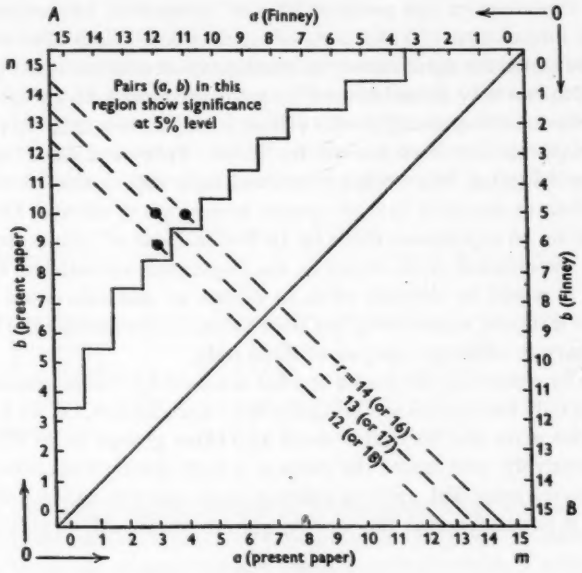
A
a (Finney)
0
15 14 13 12 11 10 9 8 7 6 5 4 3 2 1 0
n 15
Pairs (a, b) in this region show significance at 5% level
b (present paper)
b (Finney)
r = 14 (or 16)
13 (or 17)
12 (or 18)
B
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
a (present paper)
m

Fig. 8

In Finney's arrangement, when dealing with the single-tail test, $a/A \geqslant b/B$, where $A = m$ and $B = n$. Thus in Fig. 8 the scales of a and b must be reversed. His table is not entered with r, but shows for a given a the highest value of b which is just significant at the 5 % level. Thus when $A = B = 15$, for $a = 11$ he gives $b = 5$, and for $a = 12$ he gives $b = 6$. These points correspond to $r = 16$, $a = 11$ and $r = 18$, $a = 12$ in the diagram. It will be seen that on this basis an entry for the point $a = 12$, $b = 5$ on the intermediate diagonal with $r = 17$ is unnecessary and space is saved in tabulation. But there is no difference whatsoever in the basic calculations leading to Finney's table and to the special results we have given in Table 5 above.

STATISTICAL ANALYSIS OF A NON-ORTHOGONAL TRI-FACTORIAL EXPERIMENT

By W. L. STEVENS, *Universidade de São Paulo, Brazil*

1. Introduction

The formal mathematical solution to the problem of 'fitting constants' to a set of non-orthogonal data presents no great difficulty, but at first sight the arithmetical labour of solving the normal equations may appear so great as to discourage the experimenter. It is therefore important to show, by suitable worked examples, that the arithmetical procedure can be reduced to a routine so that, although the computations will always be lengthy, they can at least be seen to follow a simple and recognizable pattern.

The classical treatment of the problem by the 'method of least squares' pays scant attention to the formulation of valid tests of significance. When interactions of various factors have to be tested for significance, the mathematical solution is likely to appear very complex and much less easy to understand than the arithmetic of a worked example.

Special cases of non-orthogonality, such as those arising from a 'missing plot' or a missing row in a Latin square, have been treated by Fisher, Yates and other writers who have produced elegant solutions, but we are concerned here with a breakdown of the orthogonality restrictions so complete that no special devices are of service. The example to be discussed arose from an experiment made by Dr Rocha Faria of Lisbon, and the data were kindly placed at the disposal of the writer by the Portuguese statistician, Sr Augusto J. de Oliveira. Since it would be difficult even to invent an example more suitable for an illustration of the methods, opportunity has been taken to demonstrate the full arithmetical procedure necessary to effect an analysis of these data.

We may begin by reminding the reader of what is meant by 'fitting constants'. Suppose, for example, that only two factors are being studied—sex and diet, i.e. we have a number of animals of the two sexes and we divide them into three groups to be fed on three diets, A, B and C respectively, and record the gains in weight during a specified period. If the two sexes, even on the same diet, grow at different rates, and if in addition the diets produce different effects, it is possible that the *expected* gain in weight will be given by the formula

$$E+E_S+E_D,$$

where E is a constant, E_S is one of two constants, according to whether the animal is male or female, and E_D is one of three constants according to which of the three diets is given. The observed gain in weight will, of course, differ from the expected gain, owing to unassignable biological variation, errors of weighing, etc., but we are supposing that this 'residual' variation is random and would average zero over a sufficiently large number of animals.

The constants appearing in the above formula may be termed 'effects', since E_S measure the effect of sex and E_D measure the effects of diet. The first constant E may be called, for want of a better name, the 'general effect'. It will be noticed that they are to some extent arbitrary, for if, say, we increase every E_D by 5 units and diminish E by 5 units, the answers given by the formula will be unaltered. In fact, it is only the differences between the two values of E_S and between the three values of E_D which will be measured by an experiment.

The ambiguity could, of course, be removed by saying that the sum of the two values of E_S or of the three of E_D shall be zero, but generally we see no advantage in adopting such a convention.

It will also be noticed that the adoption of such a formula implies that the effects of sex and diet are additive, i.e. that the sex difference is the same on all diets or (equivalently) the dietary differences are the same in both sexes. Such an assumption may or may not be justified—it is possible, for example, that males respond best to one diet while females respond best to another. In general, if the effects are not additive, the formula required would be

$$E + E_{S,D},$$

where $E_{S,D}$ is one of six constants corresponding to the six combinations of two sexes and three diets.

If the first formula is adequate, the two factors, sex and diet, are said to be *independent*; if the second formula is needed, there is said to be an *interaction* between sex and diet. It is seen that the analysis of the data poses two problems:

(*a*) What type of formula is needed?

(*b*) What are the best estimates of the numerical values of the constants in this formula?

The solution of both problems is very easy when the design of the experiment ensures orthogonality between sex and diet. For orthogonality, it is not necessary that the numbers in all six classes (generated by two sexes × three diets) should be equal; it is sufficient if they obey a simple rule of proportionality illustrated in the following example:

Numbers of test animals

Diet	A	B	C	Total
Male	8 (=2×4)	10 (=2×5)	6 (=2×3)	24 (=2×12)
Female	12 (=3×4)	15 (=3×5)	9 (=3×3)	36 (=3×12)
Total	20 (=5×4)	25 (=5×5)	15 (=5×3)	60 (=5×12)

When, however, the orthogonality relation no longer holds, there is no simple method of testing the significance of interaction and of estimating the required constants. It becomes necessary to solve large sets of simultaneous linear equations which is only practicable if a routine can be developed which will converge on the solutions through a series of approximations. Furthermore, a valid analysis of variance can no longer be constructed by the standard methods with which the reader is probably familiar. When, as in the example to be discussed, there are more than two factors in the experiment, these difficulties are aggravated.

2. THE EXPERIMENT AND DATA

The experiment was designed to observe the effect on the growth of guinea-pigs of four diets distinguished by the four types of wheat which they included:

A = soft (70%), B = soft (100%), C = hard (70%), D = hard (100%).

The guinea-pigs were of two sexes and drawn from four litters, so there are three factors: sex, diet and litter.

The data are shown in Table 1, where the complete lack of orthogonality will be apparent. Although no particular interest attaches to the difference between litters, yet the effect of litter will have to be measured in order that it may be eliminated from comparisons between sexes and comparisons between diets. Thus for the purpose of analysis the experiment is tri-factorial.

It is pertinent to describe how, with the material at his disposal, the experimenter might have designed an experiment free from the disadvantages which usually attach to non-orthognality. From the guinea-pigs used in the experiment, it is possible to make up three female and seven male *triads*, i.e. sets of three of the same sex and from the same litter. A further litter would probably have provided a fourth female and an eighth male triad. From the four diets one can construct four sets of three by dropping one at a time. The sets of three diets are then allotted to the triads of animals according to the following scheme:

Litter	Males	Females
I	ABC ABC	ABC
II	BCD BCD	BCD
III and V	CDA CDA	CDA
IV	DAB DAB	DAB

Within each triad the designated three diets are allocated at random.

The experiment described (an example of *balanced incomplete blocks*) is not orthogonal but the departures from orthogonality follow a certain regular pattern, as a consequence of which the statistical analysis, while not so simple as for an orthogonal design, are nevertheless relatively straightforward and follow a standard method.

Other 'balanced' designs are available; the one illustrated was chosen because it gives information of high precision on the effects of diets, of sexes and of the interaction between diet and sex, while yielding only relatively poor information on the differences between litters.

Balanced designs have been described in order to emphasize that modern resources are quite adequate for dealing with the kind of difficulty which confronted this experimenter, and that unbalanced non-orthogonality with its attendant complications could easily have been avoided if he had consulted a statistician *before* instead of *after* doing the experiment. However, this is no answer to a research worker who has spent time and money on his experiment and has discovered too late that better methods are available. The statistician must provide a technique for analysing the data and hope that the arithmetical labours required will be sufficient to discourage the experimenter from ever again disregarding the principles of good experimental design.

3. Preliminary analyses of variance

The first stage in the computation, set out in Table 1, is almost self-explanatory. There is some replication present in the experiment, i.e. animals which belong to the same litter and sex and receive the same diet, and as the replicates are all pairs, the 'error sum of squares' may be found from the sum of squares of the differences divided by two. The sum of squares

Table 1. *Gains in weight (g.) of thirty-six guinea-pigs and first stage of the computation*

Type of wheat in diet	Litter and sex (M. or F.) I M.	I F.	II M.	II F.	III M.	III F.	IV M.	IV F.	Total M.	Total F.	Total Both
A = soft (70%)	43	58	73	—	81	62	67	71	381	191	572
	58		59						(6)	(3)	(9)
B = soft (100%)	93	60	75	71	101	76	100	—	541	207	748
	83		89						(6)	(3)	(9)
C = hard (70%)	91	70	85	70	92	—	106	73	462	271	733
				58	88				(5)	(4)	(9)
D = hard (100%)	89	—	98	69	105	—	109	76	598	217	815
	89			72	108				(6)	(3)	(9)
Total	546	188	479	340	575	138	382	220	1982	886	
	(7)	(3)	(6)	(5)	(6)	(2)	(4)	(3)	(23)	(13)	
	734		819		713		602				$X = 2868$
	(10)		(11)		(8)		(7)				(36)

Note. The numbers in brackets are the numbers of observations contributing to the respective totals.

Grand total,	$X =$	2,868
Number of observations,	$n =$	36
Sum of squares of the observations,	$S(x^2) =$	238,204
	$X^2/n =$	228,484
Sum of squares of deviations,	$S(x-\bar{x})^2 =$	9,720

Differences within the nine pairs which are alike in all respects:

A	15	—	14	—	—	—	—	—	
B	10	—	14	—	—	—	—	—	
C	—	—	—	12	4	—	—	—	
D	0	—	—	3	3	—	—	—	

Sum of squares of differences, $15^2 + 10^2 + \ldots + 3^2 = 895$
Hence entry for analysis of variance table $= 895/2 = 448$

Preliminary analysis of variance

	Degrees of freedom	Sum of squares	Mean square	Significance
Between 'combinations'	26	9272	357	***
Error (i.e. within pairs)	9	448	49·8	
Total	35	9720		

Conclusion. There is decisive evidence of differences between the twenty-seven represented combinations.

Table 2. *Condensations of the data to two-factor tables, and computation of sums of squares*

Diet	Litter I	II	III	IV	Total
A	159	132	143	138	572
	(3)	(2)	(2)	(2)	(9)
B	236	235	177	100	748
	(3)	(3)	(2)	(1)	(9)
C	161	213	180	179	733
	(2)	(3)	(2)	(2)	(9)
D	178	239	213	185	815
	(2)	(3)	(2)	(2)	(9)
Total	734	819	713	602	2868
	(10)	(11)	(8)	(7)	(36)

Diet	Sex Male	Female	Total
A	381	191	572
	(6)	(3)	(9)
B	541	207	748
	(6)	(3)	(9)
C	462	271	733
	(5)	(4)	(9)
D	598	217	815
	(6)	(3)	(9)
Total	1982	886	2868
	(23)	(13)	(36)

Sex	Litter I	II	III	IV	Total
Male	546	479	575	382	1982
	(7)	(6)	(6)	(4)	(23)
Female	188	340	138	220	886
	(3)	(5)	(2)	(3)	(13)
Total	734	819	713	602	2868
	(10)	(11)	(8)	(7)	(36)

Note. Numbers in brackets are the numbers of observations contributing to the respective subtotals.

Sums of squares:

Single factors			Pairs of factors		
			Diets	Litters	Sexes
Sexes	Diets	Litters	Litters	Sexes	Diets
231,181	232,022	230,172	234,507	232,970	235,763
228,484	228,484	228,484	228,484	228,484	228,484
2,697	3,538	1,688	6,023	4,486	7,279

Sums of squares for interactions:

Diets	3538	Litters	1688	Sexes	2697
Litters	1688	Sexes	2697	Diets	3538
Interaction (by difference)	797	Interaction	101	Interaction	1044
Diets, litters	6023	Litters, sexes	4486	Sexes, diets	7279

Table 3. *Second analysis of variance (approximate)*

Source of variation	Degrees of freedom	Sum of squares	Mean square	Variance ratio	Significance
Sex	1	2697	2697	54·2	***
Diet	3	3538	1179	23·7	***
Litter	3	1688	563	11·3	**
	7	7923	1132		
2-factor interactions:					
Diet/litter	9	797	88·6	1·78	*n.s.*
Litter/sex	3	101	37·0	0·74	*n.s.*
Sex/diet	3	1044	348	6·99	**
	15	1942			
3-factor interaction: sex/diet/litter	4	−593	(?)		
Between combinations	26	9272			
Error	9	448	49·8		
Total	35	9720			

between combinations of litter, sex and diet is then found by subtraction, thus reversing the usual procedure. Here and elsewhere the following code is used for denoting significance:

***	virtual certainty	$P < 0{\cdot}1\,\%$
**	highly significant	$0{\cdot}1\,\% < P < 1\,\%$
*	significant	$1\,\% < P < 5\,\%$
[?]	suggestive	$5\,\% < P < 20\,\%$
n.s.	not significant	$20\,\% < P$

The short analysis of variance shows that some at least of litter, sex and diet must be exerting significant effects and we can proceed to examine these in more detail.

The analysis of variance into the items corresponding to the three primary factors and their interactions is a simple extension of the method used for an ordinary orthogonal three-factor experiment. The data is condensed (Table 2) into three two-way tables showing the subtotals for all combinations of diet × litter, of sex × diet and of litter × sex. Since, as a consequence of non-orthogonality, the number of observations contributing to these subtotals will vary, the actual numbers should be inserted in brackets below the corresponding subtotals (see Table 2).

The sum of squares for any primary factor is found in the way usual when the numbers of observations are unequal, e.g. for 'between litters' we have

$$\frac{734^2}{10}+\frac{819^2}{11}+\frac{713^2}{8}+\frac{602^2}{7}-\frac{2868^2}{36}=1688.$$

Corresponding to any pair of factors, we may find a sum of squares between all combinations of the two factors. Thus between combinations of diet and litter the sum of squares is

$$\frac{159^2}{3}+\frac{132^2}{2}+\frac{143^2}{2}+\ldots+\frac{185^2}{2}-\frac{2868^2}{36}=6023.$$

From this sum we may remove the sums of squares corresponding respectively to the two single factors, diet and litter, and leave something which may be labelled 'interaction between diet and litter'.

Notice however that, as a consequence of non-orthogonality, the primary factors are not wholly independent, and the items labelled 'interaction' are therefore not true interactions nor even true 'sums of squares'. This becomes very evident when we calculate the three-factor interaction by difference (see Table 3) and find that we have run into serious trouble in respect of both degrees of freedom and 'sum of squares'. Only 4 degrees of freedom remain (although the three-factor interaction should have $3 \times 3 \times 1 = 9$ degrees of freedom) and the 'sum of squares' appears negative!

It is evident that the analysis of variance displayed in Table 3 is not genuine, but it will nevertheless prove a useful guide to decide what kind of formula will be needed to represent the effects of the various factors. If, therefore, we take the analysis of variance at its face value, it would appear that all three primary factors have significant effects, and that it is likely also that a significant interaction will be found between sex and diet.

We shall therefore begin by finding sets of 'constants' appropriate to three independent primary factors and afterwards examine the interaction between sex and diet.

4. Effects of three independent factors

If the three factors are independent, the expected value of any observation is of the form

$$E + E_S + E_D + E_L,$$

where E is a constant, E_S is one of two constants for the two sexes, E_D is one of four constants for the four diets, and E_L is one of four constants for the four litters. We do not assume that the constants within any set add to zero, and for practical purposes it is quite immaterial if, for example, every E_S is increased by 5 units and every E_D diminished correspondingly by 5 units.

The computation of the 'effects' (as we have termed these constants) is set out in Table 4. The first step (*i*) is to note, as a guide in subsequent computations, the breakdown of the numbers of observations contributing to the various marginal subtotals. Thus under 'Diets' and against A we find:

$$9 = 3 + 2 + 2 + 2 = 6 + 3,$$

meaning that there were 9 guinea-pigs on diet A and that these were drawn 3, 2, 2 and 2 from litters I, II, III and IV respectively, and that they consist of 6 males and 3 females.

The marginal subtotals are drawn from Table 2 and set out in (*ii*) of Table 4. The computational procedure may be regarded as answering the question: 'What must I add to or deduct from each observation, in respect of sex (diet, litter) in order to ensure that the means of both sexes (all diets, all litters) shall be equal?'

Thus the means for the two sexes are

Males	$1982 \div 23 = 86{\cdot}17$
Females	$886 \div 13 = 68{\cdot}15$
	$18{\cdot}02$

The average male exceeds the average female by 18·02 g. Two significant figures are sufficient at this stage, so we shall accordingly *deduct* 18 from every *male* observation. The individual adjusted observations are not required, and it is only necessary (and easier) to

Table 4. *Calculation of adjustments for three independent factors (first cycle)*

i Composition of the subtotals:

Sexes		Diets		Litters	
	(diets) (litters)		(litters) (sexes)		(sexes) (diets)
M.	23=6+6+5+6=7+6+6+4	A	9=3+2+2+2=6+3	I	10=7+3=3+3+2+2
F.	13=3+3+4+3=3+5+2+3	B	9=3+3+2+1=6+3	II	11=6+5=2+3+3+3
		C	9=2+3+2+2=5+4	III	8=6+2=2+2+2+2
		D	9=2+3+2+2=6+3	IV	7=4+3=2+1+2+2

ii Original marginal subtotals. Calculate adjustment for sex:

Subtotals	Means	Adjustments	Subtotals	Subtotals
1982	86·17	−18	572	734
886	68·15	0	748	819
			733	713
			815	602
2868			2868	2868

iii Adjust all subtotals and calculate adjustments for diet:

Subtotals	Sub-totals	Means	Adjust-ments	Subtotals
1568	464	51·56	+20	608
886	640	71·11	0	711
	643	71·44	0	605
	707	78·56	−7	530
2454	2454			2454

iv Adjust all subtotals and calculate adjustments for litter:

Subtotals	Subtotals	Sub-totals	Means	Adjust-ments
1646	644	654	65·40	+14
925	640	730	66·36	+13
	643	631	78·88	0
	644	556	79·43	0
2571	2571	2571		

Table 4 (*continued*): *Second cycle and final adjustments*

Sexes			Diets			Litters		
Subtotals	Means	Adjustments	Sub-totals	Means	Adjust-ments	Sub-totals	Means	Adjust-ments
v Adjust subtotals for litter and calculate adjustments for sex:								
1822	79·22	+0·16	712			794		
1032	79·38	0	721			873		
			710			631		
			711			556		
2854			2854			2854		
vi Adjust subtotals for sex and calculate adjustments for diet:								
1825·68			712·96	79·22	+1·00	795·12		
1032			721·96	80·22	0	873·96		
			710·80	78·98	+1·24	631·96		
			711·96	79·11	+1·11	556·64		
2857·68			2857·68			2857·68		
vii Adjust subtotals for diet and calculate adjustments for litter:								
1844·54			721·96			802·82	80·28	−0·28
1043·29			721·96			883·01	80·27	−0·27
			721·96			638·66	79·83	+0·17
			721·95			563·34	80·48	−0·48
2887·83			2887·83			2887·83		
viii Adjust subtotals for litter and calculate final adjustments for all factors:								
1840·06	80·003	−0·003	719·96	79·996	0	800·02	80·002	0
1040·00	80·000	0	720·17	80·019	−0·023	880·04	80·004	−0·002
			719·97	79·997	−0·001	640·02	80·002	0
			719·96	79·996	0	559·98	79·997	+0·005
2880·06			2880·06			2880·06		

ix Adjust grand total and calculate general effect:

Grand total (from *viii*)	2880·06
Adjustment for sex	−0·069
Adjustment for diet	−0·216
Adjustment for litter	+0·013
Final grand total	2879·788
General effect (i.e. mean)	79·99411

Table 5. *Collection of adjustments and computation of sum of squares*

Factor	First cycle	Second cycle	Final (*viii*)	Total adjustment	Effect	Subtotals (*ii*)
Sex M.	(*ii*) −18	(*v*) +0·16	−0·003	−17·843	+17·843	1982
F.	0	0	0	0	0	886
Diet A	(*iii*) +20	(*vi*) +1·00	0	+21·000	−21·000	572
B	0	0	−0·023	− 0·023	+ 0·023	748
C	0	+1·24	−0·001	+ 1·239	− 1·239	733
D	− 7	+1·11	0	− 5·890	+ 5·890	815
Litter I	(*iv*) +14	(*vii*) −0·28	0	+13·720	− 13·720	734
II	+13	−0·27	−0·002	+12·728	−12·728	819
III	0	+0·17	0	+ 0·170	− 0·170	713
IV	0	−0·48	+0·005	− 0·475	+ 0·475	602
Note. The small roman numerals indicate from where the figures have been drawn				(*ix*)	+79·99411 General effect	2868 Grand total

Sum of products of effects and corresponding subtotals = 236,355
X^2/n = 228,484
Sum of squares for effects of factors = 7,871

Table 6. *Third analysis of variance*

Source of variation	Degrees of freedom	Sum of squares	Mean square	Variance ratio	Significance
Effects of the 3 factors	7	7871	1124	22·5	***
Interactions	19	1401	73·7	1·48	*n.s.*
Between 'combinations'	26	9272			
Residual	9	448	49·8		
Total	35	9720			

record the effects of these adjustments on all the subtotals. Thus the new male subtotal will be

$$1982 - 23(18) = 1568.$$

The female subtotal will remain unaltered. The subtotal for diet A is based on 6 males (and 3 females) as is shown in (*i*) of Table 4, and will therefore become

$$572 - 6(18) = 464.$$

Similarly, the subtotal for litter I is based on 7 males (and 3 females) and will therefore become

$$734 - 7(18) = 608.$$

The adjusted subtotals are set out in (*iii*) of Table 4. It is noted that agreement between the totals of the three sets provides a complete check on the adjustment of the subtotals.

Next the four diet means are computed:

$$\cdot 51{\cdot}56 \quad 71{\cdot}11 \quad 71{\cdot}44 \quad 78{\cdot}56.$$

Since two significant figures are sufficient it will be convenient to make no adjustment for diets B and C, add 20 for diet A and subtract 7 for diet D:

$$51{\cdot}56 + 20 = 71{\cdot}56, \quad 71{\cdot}11, \quad 71{\cdot}44, \quad 78{\cdot}56 - 7 = 71{\cdot}56.$$

These adjustments are carried through into all subtotals as shown in (*iv*) of Table 4. Thus the new male subtotal will be

$$1568 + 6(20) + 6(0) + 5(0) + 6(-7) = 1646.$$

The means for the four litters and hence the adjustments in respect of litter are now computed, as in (*iv*) of Table 4, and the adjustments made in (*v*). This completes the first cycle, adjustments having been made in turn for

sex—diet—litter.

The cycle of adjustments (carried to two decimals) may now be run through a second time as set out in (*v*), (*vi*) and (*vii*) of Table 4.

The process described will converge fairly rapidly to a solution, i.e. we shall find that after a few cycles the adjustments have become negligibly small. However, we can judge from the drop in the magnitude of the adjustments from the first cycle to the second that the third cycle adjustments will all be small. The third cycle can therefore be abridged, as in (*viii*) of Table 4, by calculating all means and all final adjustments, to three decimals, at once.

The subtotals, modified by these final adjustments, are not required, but the grand total should be adjusted as in (*ix*) of Table 4. The mean, found on dividing this final grand total by $n = 36$, will be the estimate of E, the so-called 'general effect'. It should be calculated to one or two decimals more than are provided in the final adjustments.*

In carrying out the adjustments as just described, it is, of course, immaterial on which factor we start or in which order we pass round the cycle, but other things being equal, it will usually be best to start with the factor producing the biggest effects and work down to the one with least effect. It is also immaterial what mean we aim at when applying adjustments, but the arithmetic is generally simpler if we aim at either the largest or smallest of the set so that one of the adjustments (in each set) is zero while the others are all of the same sign. However, in the first adjustment for diet ((*iii*) of Table 4) advantage was taken of the fact that the two *intermediate* values were by chance approximately the same (71·11 and 71·44), and *two* adjustments were thus made zero with a resultant slight simplification of the arithmetic.

The adjustments should now be collected together. For clarity this process has been set out in Table 5, but usually they could be picked up and added straight from Table 4. The resultant total adjustments, *with their signs changed*, will then be estimates of the 'effects', i.e. of the constants E_S, E_D and E_L. They may be reckoned to be correct or nearly correct to three decimals.

* The reason is that the sum of squares to be subsequently computed is near its maximum value (the residual sum of squares being near a minimum). The sum of squares is therefore relatively insensitive to errors in the estimated constants. Hence even if the estimates are good, say, to only three decimals, they may be treated, *for the purpose of calculating a sum of squares*, as though they were good to four or five decimals.

In the last column of Table 5 are written the *original* marginal subtotals against the corresponding effects and, at the bottom, the *original* grand total against the *final* adjusted mean (i.e. general effect). The sum of squares for the effects of the three factors is found by adding the products of the last two columns, not omitting the general effect × grand total, and subtracting the usual term, $X^2/n = (2868)^2/36$:

$$(17{\cdot}843)\,(1982)+(0)\,(886)+\ldots+(79{\cdot}99411)\,(2868)-228{,}484=7871.$$

The number of degrees of freedom to be attributed to this sum of squares is 7, for, although 10 effects have been estimated, any arbitrary constant could be added to the effects in any one of the 3 sets (sex, diet and litter) and taken off the general effect without affecting the result, whence the number of degrees of freedom is

$$10-3=7.$$

Notice that sums of squares are not obtained individually for the three factors but only one sum of squares for the three acting together. The sum of squares, 7871, is not greatly different from that obtained in the approximate analysis of variance, 7923 (see Table 3), which suggests that the first four lines of the approximate analysis are all fairly reliable.

Subtracting the sum of squares for the effects (with 7 degrees of freedom) from the sum of squares between all combinations of the 3 factors (26 degrees of freedom) we are left with a sum of squares with 19 degrees of freedom representing various undifferentiated interactions between the factors (see Table 6). The corresponding mean square is larger, but not significantly larger, than the residual mean square, and hence there is at first sight no good evidence for interaction. However, the approximate analysis of variance (Table 3) did suggest that there is a significant interaction between sex and diet, and it is conceivable that this interaction, with only 3 degrees of freedom, has been hidden by being diluted with other interactions. It is therefore proper to make an exact analysis to test the significance of the sex × diet interaction.

5. THREE FACTORS WITH INTERACTION BETWEEN TWO OF THEM

If no restrictions are placed on the form of the interaction between sex and diet, then there will be an 'effect' appropriate to *every* combination of these two factors. This means that together they constitute virtually a single factor whose 'levels' are specified by the combination, e.g. male on diet A, female on diet C, etc. The data can therefore be treated as though they had arisen from a bi-factorial experiment.

The procedure for calculating the adjustments, as set out in Table 7, does not differ greatly from that employed with three factors. Sex × diet is now a single factor, and it is only to save paper that the 8 combinations have been laid out in a 4×2 table. It is found convenient to display the 4 litters horizontally instead of vertically, so that the block of numbers under litter I, etc., in (*i*) of Table 7 have the right pattern and orientation for calculating the effect of the sex × diet adjustments on the litter subtotal. Thus having found the adjustments for the eight sex × diet combinations in (*ii*) of Table 7, we calculate the adjusted subtotal for litter I as follows:

$$734+\begin{matrix} 2(0) & 1(0) \\ -2(27) & -1(5) \\ -1(29) & -1(4) \\ -2(36) & 0(9) \end{matrix}=570$$

The convergence is not so rapid as in Table 4, and the third decimal is therefore of doubtful value.

The sum of squares for effects (calculated in Table 8) has 10 degrees of freedom since there are now *two* sets of effects, one with 8 and one with 4, and

$$8+4-2=10.$$

Alternatively we may argue that there are 7 independent comparisons between the 8 sex × diet combinations and 3 independent comparisons between the 4 litters. Hence number of degrees of freedom is

$$7+3=10 \quad \text{as before.}$$

The sum of squares (7 degrees of freedom) for the three independent factors previously obtained may now be subtracted from this sum of squares with 10 degrees of freedom to leave a sum of squares with 3 degrees of freedom representing the interaction between sex and diet, as set out in Table 9. The balance of 16 degrees of freedom to make up the 26 degrees of freedom between all combinations represents undifferentiated interactions between sex and litter, diet and litter and between all three factors.

From the analysis of variance (Table 9) it appears that not only are the other interactions non-significant, but their mean square is scarcely greater than the residual mean square. The variance ratio for the sex × diet interaction is

$$F = 196{\cdot}0/49{\cdot}8 = 3{\cdot}936,$$

and is therefore just significant at the 5 % level. There is accordingly some moderately good evidence of an interaction between sex and diet.

We may now examine the nature of this interaction by listing the differences, male *minus* female, on the four diets:

Diet	Effects (from Table 8)		Difference M. – F.
	Male	Female	
A	+ 3·090	0	+ 3·090
B	+29·757	+ 8·933	+20·824
C	+29·100	+ 7·590	+21·510
D	+37·800	+10·353	+27·447

The table gives the impression that diet A may be the exception. For A the responses of the two sexes are not greatly different, while for B, C and D there is a big difference between the two sexes, but this difference may possibly not vary significantly among the three diets. On the other hand, the data also suggest the alternative hypothesis that the degree of difference between the sexes rises with the response, i.e. that the better the diet, the more the males will outstrip the females. It is very necessary to remember that the same data may be consistent with several hypotheses which, biologically considered, are fundamentally different. It is legitimate, and indeed desirable, to employ *a priori* biological knowledge at this stage even though the knowledge may amount to little more than a hint that one hypothesis may be more 'reasonable' than another. Often, however, no *a priori*

Table 7. *Calculation of adjustments for two factors, sex × diet and litter (first two cycles)*

i Composition of the subtotals:

Sex × diet combinations			Litters			
	Male	Female	I	II	III	IV
A	6=2+2+1+1	3=1+0+1+1	10	11	8	7
B	6=2+2+1+1	3=1+1+1+0	=	=	=	=
			2+1	2+0	1+1	1+1
C	5=1+1+2+1	4=1+2+0+1	2+1	2+1	1+1	1+0
			1+1	1+2	2+0	1+1
D	6=3+1+2+1	3=0+2+0+1	2+0	1+2	2+0	1+1

ii Original marginal subtotals. Calculate adjustments for sex × diet:

Subtotals		Means		Adjustments		Subtotals			
381	191	63·50	63·67	0	0	734	819	713	602
541	207	90·17	69·00	−27	−5				
462	271	92·40	67·75	−29	−4				
598	217	99·67	72·33	−36	−9				
2868				Totals		2868			

iii Adjust all subtotals and calculate adjustments for litter:

381	191	Subtotals	570	669	551	497
379	192	Means	57·00	60·82	68·88	71·00
317	255	Adjustments	+14	+10	+2	0
382	190					
2287		Totals	2287			

iv Adjust all subtotals and calculate adjustments for sex × diet:

431	207	71·83	69·00	−2·83	0	710	779	567	497
429	218	71·50	72·67	−2·50	−3·67				
345	289	69·00	72·25	0	−3·25				
424	210	70·67	70·00	−1·67	−1·00				
2553				Totals		2553			

v Adjust all subtotals and calculate adjustments for litter:

414·02	207·00	Subtotals	689·08	754·50	554·66	485·75
414·00	206·99	Means	68·91	68·59	69·33	69·39
345·00	276·00	Adjustments	+0·48	+0·80	+0·06	0
413·98	207·00					
2483·99		Totals	2483·99			

Table 7 (*continued*): *Final adjustments*

vi Adjust all subtotals and calculate final adjustments for both factors:

Sex × diet combinations						Litters			
416·64	207·54	69·440	69·180	−0·260	0	693·88	763·30	555·14	485·75
416·62	208·33	69·437	69·443	−0·257	−0·263	69·388	69·391	69·392	69·393
346·40	278·08	69·280	69·520	−0·100	−0·340	+0·005	+0·002	+0·001	0
415·86	208·60	69·310	69·533	−0·130	−0·353				
2498·07			Totals			2498·07			

vii Adjust grand total and calculate general effect:

Grand total (from *vi*)	= 2498·07
Adjustments for sex × diet	= −7·590
Adjustments for litter	= +0·080
Final grand total	= 2490·560
General effect (i.e. mean)	= 69·18222

Table 8. *Collection of adjustments and calculation of sum of squares*

Factor		First cycle	Second cycle	Final (*vi*)	Total adjustment	Effect	Original subtotals
Combinations of sex and diet	M., A	(*ii*) 0	(*iv*) −2·83	−0·260	−3·090	+3·090	381
	M., B	−27	−2·50	−0·257	−29·757	+29·757	541
	M., C	−29	0	−0·100	−29·100	+29·100	462
	M., D	−36	−1·67	−0·130	−37·800	+37·800	598
	F., A	0	0	0	0	0	191
	F., B	−5	−3·67	−0·263	−8·933	+8·933	207
	F., C	−4	−3·25	−0·340	−7·590	+7·590	271
	F., D	−9	−1·00	−0·353	−10·353	+10·353	217
Litters	I	(*iii*) +14	(*v*) +0·48	+0·005	+14·485	−14·485	734
	II	+10	+0·80	+0·002	+10·802	−10·802	819
	III	+2	+0·06	+0·001	+2·061	−2·061	713
	IV	0	0	0	0	0	602
General effect and grand total (*vii*)						+69·18222	2868

Sum of products of effects and corresponding subtotals	= 236,343
X^2/n	= 228,484
Sum of squares for effects of factors	= 8,459

Table 9. *Fourth analysis of variance*

Source of variation	Degrees of freedom	Sum of squares	Mean square	Variance ratio	Significance
Three independent factors	7	7871	1124	22·6	***
Sex × diet interaction	3	588	196·0	3·936	*
Three factors and all sex × diet interactions	10	8459			
Other interactions	16	813	50·8		n.s.
Between all combinations	26	9272			
Residual	9	448	49·8		
Total	35	9720			

knowledge is relevant, and in such cases the greatest care should be taken to avoid adopting an hypothesis merely because it is not belied by the data.

It is, of course, legitimate to test whether the data do or do not contradict *any hypothesis*, always remembering that failure to contradict the hypothesis is not a proof that the hypothesis is true. Hence, although it is not necessarily true that the most appropriate hypothesis to test is that diet A is exceptional, there can be no objection to making such a test as an illustration of the statistical technique.

The hypothesis to be examined is equivalent to supposing that one particular degree of freedom out of the three degrees of freedom for sex × diet interaction has a real effect while the other two have no effect. Hence we should separate the mean squares for the one and the two degrees of freedom respectively and endeavour to show that the former is and the latter is not significant.

Any algebraic formulation of the process would be cumbersome, but the arithmetic proceeds quite straightforwardly once the pattern of adjustments is appreciated (see Table 10). We are now permitting a different sex difference in diet A on the one hand and in diets B, C and D on the other. Hence the adjustment for sex is made separately in the two groups of diets (A) and (B, C or D). To avoid bandying adjustments backwards and forwards between sexes and diets, it is advisable to adjust only the male results. The means of the two sexes on diet A only are

Male	381/6 = 63·50
Female	191/3 = 63·67

These are sufficiently close to make no adjustment necessary on the first cycle. The sex means calculated from the pooled data for diets B, C and D are

Male	1601/17 = 94·18
Female	695/10 = 69·50
Difference	= 24·68

Hence adjustment is made by subtracting 25 from every male on diets B, C or D, leaving unaltered both males and females on diet A and females on diets B, C and D.

The adjustments to the subtotals are made in (*iii*) of Table 10. Thus for diets

$$\text{A} \quad 572 - 6(0) = 572 \text{ etc.}$$

and for litters

$$\text{I} \quad 734 - 5(25) = 609 \text{ etc.}$$

The final adjustments shown in (*viii*) of Table 10 are appreciable, so the third decimal is of doubtful value.

The number of degrees of freedom in the sum of squares for effects is now 8, being made up of 1 for sex in diet A, 1 for sex in diets B, C and D, 3 for differences between diets and 3 for differences between litters:

$$1 + 1 + 3 + 3 = 8.$$

The sum of squares with 7 degrees of freedom obtained with three independent factors (see Table 6) may be subtracted from the sum of squares with 8 degrees of freedom, to leave a square of 1 degree of freedom which measures the effect of the specific interaction which we have selected for examination. The balance to bring the sum of squares with 8 degrees of freedom up to the sum of squares with 10 degrees of freedom found when all interactions between sex and diet were permitted (see Table 9) is a sum of squares with 2 degrees of freedom and measures the effect of the interaction between sex and diet *other than* the specific degree of freedom which we selected. The full analysis of variance is set out in Table 12.

It is clear that the hypothesis examined is acceptable on these data. The selected degree of freedom in interaction is fairly significant, whereas the mean square for the remaining 2 degrees of freedom is neither significant nor even much greater than the residual sum of squares. In judging the significance of the selected degree of freedom one should, of course, allow for the fact that it was chosen as being the biggest of three. However, the three together were shown to be on the border-line of significance, so there is little danger that we have misled ourselves by selecting a square which is large only by chance.

6. Presentation of conclusions

Since litter is independent of the other two factors, any comparison between sexes, between diets or between any combinations of sex and diet may be regarded as generally true of all the litters in this experiment.

There is some evidence that sex and diet are not independent factors, i.e. that the two sexes react differentially to the diets. This evidence only just passes the conventional 5% level of significance, so the possibility must not be ruled out that an interaction of the observed magnitude has arisen fortuitously. It is impossible, on statistical evidence alone, to specify at all closely the nature of this interaction, for it is consistent with either of the following hypotheses or indeed with many others:

(*a*) That the difference between male and female is the same on diets B, C and D but different on A.

(*b*) That the sex difference increases steadily with the response of either sex, i.e. that the 'better' the diet, the more the male will outstrip the female.

If the decision to test either of these hypotheses had been made before examining the data, the significance of the particular degree of freedom examined would have been considered reasonably strong. Of course when, as in the present case, the hypothesis is suggested by the data themselves, one must be prepared to discount to some extent the impression given by the moderately high level of significance.

Table 10. *Calculation of adjustments for three factors with one degree of freedom in interaction*

i Composition of subtotals:

Sexes × diets	Diets	Litters
Diet A (litters) M. 6= 6 =2+2+1+1 F. 3= 3 =1+0+1+1	(litters) (sexes) A 9=3+2+2+2=6+3	(diets) I 10=3+3+2+2 2+ 5 =7 males
	B 9=3+3+2+1=6+3	II 11=2+3+3+3 2 +4 =6 males
Diets B+C+D M. 17=6+5+6=5+4+5+3 F. 10=3+4+3=2+5+1+2	C 9=2+3+2+2=5+4	III 8=2+2+2+2 1 +5 =6 males
	D 9=2+3+2+2=6+3	IV 7=2+1+2+2 1 +3 =4 males

ii Original marginal subtotals. Calculate adjustments for sex × diet:

Sub-totals	Means	Adjust-ments	Sub-totals	Means	Adjust-ments	Sub-totals	Means	Adjust-ments
381	63·50	0	572			734		
191	63·67	0	748			819		
1601	94·18	−25	733			713		
695	69·50	0	815			602		
2868			2868			2868		

iii Adjust all subtotals and calculate adjustments for diet:

381			572	63·56	+10	609		
191			598	66·44	+7	719		
1176			608	67·56	+6	588		
695			665	73·89	0	527		
2443			2443			2443		

iv Adjust all subtotals and calculate adjustments for litter:

441			662			672	67·20	+14
221			661			778	70·73	+10
1248			662			634	79·25	+2
740			665			566	80·86	
2650			2650			2650		

Table 10 (*continued*)

v Adjust all subtotals and calculate adjustments for sex × diet:

Sub-totals	Means	Adjust-ments	Sub-totals	Means	Adjust-ments	Sub-totals	Means	Adjust-ments
491	81·83	−2·83	728			812		
237	79·00	0	737			888		
1368	80·47	+1·53	724			650		
820	82·00	0	727			566		
2916			2916			2916		

vi Adjust subtotals for sex × diet and calculate adjustments for diet:

474·02			711·02	79·00	+3·91	813·99		
237·00			746·18	82·91	0	888·46		
1394·01			731·65	81·29	+1·62	654·82		
820·00			736·18	81·80	+1·11	567·76		
2925·03			2925·03			2925·03		

vii Adjust all subtotals and calculate adjustments for litter:

497·48			746·21			831·18	83·12	+0·39
248·73			746·18			904·47	82·22	+1·29
1408·77			746·23			668·10	83·51	0
829·81			746·17			581·04	83·01	+0·50
2984·79			2984·79			2984·79		

viii Adjust all subtotals and calculate final adjustments for all factors:

501·34	83·557	−0·350	750·96	83·440	+0·102	835·08	83·508	+0·007
249·62	83·207	0	751·72	83·524	+0·018	918·66	83·515	0
1417·38	83·375	+0·429	751·88	83·542	0	668·10	83·512	+0·003
838·04	83·804	0	751·82	83·536	+0·006	584·54	83·506	+0·009
3006·38			3006·38			3006·38		

ix Adjust grand total and calculate general effect:

Grand total (from *viii*)	3006·38
Adjustment for sex × diet	+5·193
Adjustment for diet	+1·134
Adjustment for litter	+0·157
Final grand total	3012·864
General effect (i.e. mean)	83·69067

Table 11. *Collection of adjustments and calculation of sum of squares*

Factor	First cycle	Second cycle	Final (*viii*)	Total adjustment	Effect	Subtotals (*ii*)
Sex in diet A						
M.	(*ii*) 0	(*v*) −2·83	−0·350	−3·180	+3·180	381
F.	0	0	0	0	0	191
Sex in diets B, C and D						
M.	−25	+1·53	+0·429	−23·041	+23·041	1601
F.	0	0	0	0	0	695
Diet A	(*iii*) +10	(*vi*) +3·91	+0·102	+14·012	−14·012	572
B	+7	0	+0·018	+7·018	−7·018	748
C	+6	+1·62	0	+7·620	−7·620	733
D	0	+1·11	+0·006	+1·116	−1·116	815
Litter I	(*iv*) +14	(*vii*) +0·39	+0·007	+14·397	−14·397	734
II	+10	+1·29	0	+11·290	−11·290	819
III	+2	0	+0·003	+2·003	−2·003	713
IV	0	+0·50	+0·009	+0·509	−0·509	602
			General effect (*ix*)		+83·69067	
			Grand total (original)			2868

Sum of products of effects and corresponding subtotals = 236,817
X^2/n = 228,484
Sum of squares for effects of factors = 8,333

Table 12. *Fifth and final analysis of variance*

Reference	Source of variation	Degrees of freedom	Sum of squares	Mean square	Variance ratio
T. 6	Three independent factors	7	7871	1124	22·6***
Dif.	Selected sex × diet interaction	1	462	462	9·3*
T. 11	3 factors and a selected interaction	8	8333		
Dif.	Remaining sex × diet interactions	2	126	63	*n.s.*
T. 9	3 factors and sex × diet interactions	10	8459		
T. 9	Other interactions	16	813	50·8	*n.s.*
T. 9	Between all factorial combinations	26	9272		
T. 9	Residual	9	448	49·8	
	Total	35	9720		

If we accept the evidence for interaction between sex and diet, the results of the experiment may be summarized by the estimated effects shown in Table 8. The magnitude of the differences between litters are, however, of little interest. Each combination of sex and diet may therefore be averaged over the four litters. We have

$$69{\cdot}182 - \tfrac{1}{4}(14{\cdot}485 + 10{\cdot}802 + 2{\cdot}061 + 0) = 62{\cdot}345.$$

To this we may add the effect for each diet × sex combination; thus for males on diet A we have

$$62{\cdot}345 + 3{\cdot}090 = 65{\cdot}435.$$

It is unnecessary to quote the decimal figures; the calculation was carried to such accuracy only to illustrate the arithmetical processes. The conclusions can therefore be summarized in the following table:

Expected gains in weight (g.)

Diet	Male	Female	M. – F.
A	65	62	3
B	92	71	21
C	91	70	21
D	100	73	27

This table, read in conjunction with the notes on the significance of the sex × diet interaction, summarizes all the information in the experiment.

7. Discussion of methods

We have shown that even when orthogonality restrictions have completely broken down in a factorial experiment, it is still possible to construct exact analyses of variance and to make valid tests of the significance of any effects or interactions or in fact of any form of departure from any hypothesis. The reader may, however, have been left with an impression that the whole thing is a *tour de force* rather than a recognizable method of analysis. We regret if such is the impression; the calculations are long and tedious, but we hope that they follow a pattern of which the logic can be appreciated.

The principle of construction of a valid analysis of variance, in so far as it can be summed up in a few sentences, is as follows: Take any two hypotheses X and Y, of which X is admitted and Y is more complicated. Calculate the effects which are specified by X and the sum of squares for these effects. Do the same for hypothesis Y. Then the increases in the number of degrees of freedom and in the sum of squares as we pass from X to Y will lead to a mean square for deciding whether we must adopt the more complicated hypothesis. Thus in constructing the analysis of variance in Table 9 the relevant hypotheses were:

		Degrees of freedom
X	Sex, diet and litter have significant effects but none of their interactions is significant	7
	Difference	3
Y	Same as X, except that possibility of interaction between sex and diet is admitted	10

The difficulty which is peculiar to a non-orthogonal experiment is that the effects of irrelevant factors, such as litter in the example quoted, have to be computed separately on

the two hypotheses, X and Y. In an orthogonal experiment the effects of litters (and hence the corresponding sum of squares) would be unaltered by any modification of the hypothesis concerning the other two factors. This, when translated into general terms, explains why the statistical analysis of a non-orthogonal (and unbalanced) experiment will necessarily be troublesome; in fact, as we have already emphasized, the moral to be pointed is that the research worker should either acquaint himself with the principles of experimental design or else employ a statistician before embarking on the experiment.

SUMMARY

Data are presented from a biological experiment in which no attempt was made to ensure orthogonality or even balanced non-orthogonality. The proper design for an experiment with the same material is briefly discussed. The methods for the statistical analysis of the data are then explained by means of the arithmetical computations laid out in such a form that the routine can be understood and hence applied to other examples.

COMPARISONS OF HEIGHTS AND WEIGHTS OF GERMAN CIVILIANS RECORDED IN 1946–7 AND ROYAL AIR FORCE AND OTHER BRITISH SERIES

By G. M. MORANT, D.Sc., *R.A.F. Institute of Aviation Medicine*

1. Summary and main conclusions

1·1. Records of heights and weights collected during the war show clearly that changes were taking place in the distributions of the measurements for the British population. Compared with pre-war years maximum height, indicating skeletal maturity, was normally reached at a younger age than previously. Weights for the war period tended to be above pre-war levels for ages under about 25 years and below them for ages over 25. New standards were needed which could be used to assess the significance of contemporary and future changes in heights and weights. This paper is a contribution to that topic, which is aided by comparison of the recent British records with those for large numbers of civilians in the British zone of Germany measured for the Public Health Branch of the Control Commission for Germany (British Element) from December 1946 to June 1947.

1·2. Comparisons made lead to the following general conclusions:

(*a*) Appreciable differences are found between the mean heights, recorded weights and weights standardized for height of the German children of different towns (Hamburg and Berlin) and regions (*Länder*). For adults the only relation between the mean weights standardized for height which is consistent for both sexes and nearly all age groups is that the Berlin series are inferior to all others.

(*b*) Pooled series representing all regions except Berlin are treated in later comparisons. The data were subdivided to represent earlier (December 1946–February 1947) and later (March–June 1947) periods. None of the secular differences between the mean heights and weights of the German children, and between the weights of the adults, are large and it is possible that they were due partly to normal seasonal fluctuations which may be exhibited by a population having constant nutritional conditions.

(*c*) Occupational series of adults show a sequence in mean weights from very heavy workers (heaviest) to normal consumers (lightest).

(*d*) The total series, excluding Berlin children and adolescents, give growth curves for height which are of the usual forms, but which have maxima at the remarkably young ages of 19½ years for males and 17 years for females. For both sexes the weight curves are peculiar in showing less increase for ages over 18 than is normally found.

(*e*) For ages under 19 and comparisons of weights reduced at each age to a constant height, the post-war German standards occupy low positions, but ones which are not extremely low compared with those for pre-war German series.

(*f*) The maximum of the male age curve for height given by the post-war survey is greater than any of the pre-war means for regional sections of the male adult population of the British zone of Germany. These pre-war standards were probably underestimates of the mean heights of German men.

(*g*) The following conclusions ((*g*)–(*j*)) are derived from comparisons of mean weights for different series reduced to the same height at each age, i.e. for weights compared after

allowance has been made for differences in height. For adolescent male British industrial workers, series measured during the war (1943) were decidedly heavier than series measured before the war (1929–32). The British series of both dates are well above the post-war German series. For British female industrial workers, series measured during the war show a higher standard than pre-war series (1926), the latter being rather lower than the post-war German series.

(*h*) For men aged 17–40 years the mean standardized weights for twelve British and the post-war German series give the sequence: Commonwealth and British serving aircrew (heaviest)—aircrew recruits—R.A.F. ground staff—industrial workers, 1943—industrial workers, 1929–32—Germans—conscripts, including those rejected after examination, 1917–18—unemployed industrial workers, 1929–32 (lightest).

(*i*) The situation is markedly different for ages over 40. The German series clearly falls to the lowest place, and unlike all other series it shows mean weights rapidly declining with increasing age: the British wartime industrial workers fall below the pre-war employed workers, though remaining above the low levels of the 1917–18 conscripts and pre-war unemployed workers. Rationing has affected most markedly the weights of the middle aged and aged.

(*j*) The few series of women lead to similar conclusions. British industrial workers were decidedly heavier during than before the war, and the Germans were clearly at a lower level for all except younger ages (up to about 30 years). The evidence suggests that for all ages loss of weight due to restricted rations has been greater for females than for males, and decidedly greater in the case of both sexes for middle-aged and old people than for young adults and children.

2. THE GERMAN 1946–7 SURVEY

2·1. Starting in December 1946, the Public Health Branch of the Control Commission for Germany (British Element) has carried out a new survey of the body weights of German people in all parts of the British zone, including the British sector of Berlin. The plan was to remeasure the same people, as far as possible, at monthly intervals, the weights being taken during the first two weeks of each month. They were recorded under the direction of Public Health Officers, who aimed, in accordance with guidance received, at obtaining random samples of the populations of their districts. Both sexes and all age groups were to be represented. The returns were reduced by No. 1 Nutrition Survey Team and Hollerith machines were used for the purpose. The writer is indebted to Brigadier W. S. Martin, M.C., Public Health Adviser, Health Branch, C.C.G., and to F. D. G. Bailey, Esq., formerly Nutrition Officer of the same Branch, for permission to use the German data given in a new form in the present paper.

2·2. The records available consist of a preliminary report on the survey (Control Commission for Germany, 1947) with tabulated data for the months December 1946–March 1947, and additional tables for the months April–June 1947. Only average values of the measurements—heights and weights of children and adolescents and weights standardized for height of adults—are given. The total material can be divided into two sets, viz.

(i) Children and adolescents: (*a*) means of heights and weights as recorded for each month December 1946–February 1947 (and for some regions also for March 1947), distinguishing sexes, ages (denoted by year of birth), and Berlin, Hamburg and regional samples, the data

also being given for the total British zone except Berlin; (*b*) the same data for each month March–June 1947 for the total British zone except Berlin only.

(ii) Adults: (*a*) mean weights reduced to constant heights (see §3·6) for each month December 1946–February 1947 (and for some regions also for March 1947), distinguishing sexes, ages (by decennial groups of years of birth), occupational groups, and Berlin, Hamburg and regional samples, the data also being given for the total British zone except Berlin; (*b*) the same data for each month March–June 1947 for the total British zone except Berlin only.

The age ranges for the children and adolescent series, on the one hand, and for the 'adult', on the other, overlap. The youngest group for the latter is 7–17 years and data for this were not used as it covers too large a part of the age cycle when growth is most rapid. Other details regarding the survey are discussed in § 3 below.

3. Treatment of the German 1946–7 survey data

3·1. The available statistical data for the German survey consist of means only. Possible ways of treating them are thus very limited and rigorous comparisons cannot be made. The aim of the comparisons in following sections of this paper is to reveal the salient characteristics of the material only. The treatment consists chiefly in deriving clear-cut conclusions from graphical comparisons. The means are for series distinguished on account of sex, age, locality, time (month of measurement) and occupation (adults only). By appropriate grouping of the means, providing pooled means, the situation is examined for each of the last three of these factors considered singly, possible interaction between them being ignored. This omission is not likely to have given invalid conclusions, since notice is only taken of marked distinctions.

3·2. The aim was to remeasure the same people at monthly intervals, so that information would be obtained most economically regarding a secular change in the distributions of weights for the populations sampled. For various reasons this was not achieved fully and it is said: 'The number of subjects who were actually weighed in consecutive months is much lower than might be expected.' The number of individuals giving a mean for any particular month is known. In pooling data for different months the number of observations is known but not the number of individuals involved, because the number of repeated weighings is unknown. It may be noted that the n's for means given in the tables below have greater statistical value in secular comparisons and lesser value in regional and occupational comparisons.

3·3. The original grouping is by calendar years of birth, not years of age. When translated into years of age, with regard to the date or dates of measurement, the central values of groups are fractions of years. In comparing the data for children with other series it is more convenient to have means for even central values of the form $x{\cdot}0$ or $x{\cdot}5$ years. Estimates at these ages were obtained for the German and other series when necessary by linear interpolation, which can be supposed sufficiently precise when data at 1-year intervals are being treated. Consideration of the following points is necessary.

3·4. *Seasonal fluctuations in height and weight increments.* The German data are for the months December 1946–June 1947, measurements having been recorded during the first two weeks of each month. It is convenient to make comparisons between pooled means for

the two periods December 1946–February 1947 and March–June 1947. In doing this, differences observed might be partly due to normal seasonal changes in growth rates for height and weight in the case of immature individuals, and to normal seasonal fluctuations in the weights of adults. The same question has to be taken into account in comparing the German series as a whole with others. Data regarding seasonal changes of heights and weights in children have been given by Schmid-Monnard (1895), Mumford (1927), Friend (1935) and Friend & Bransby (1947), and for adolescents and adults by Kemsley (1945–7). It has been found for schoolchildren that growth at all times of the year is greater during holidays than during term time. The evidence of different series is not entirely in agreement, but it justifies some rather vague general conclusions. These are that seasonal fluctuations in the growth rate of height in boys (and presumably girls, too) are probably so small that they may be neglected in comparing different series; that they may be appreciable for weights of children—giving maximum differences between quarterly averages for the same population of the order 2 lb.—and that they are probably negligible for weights of adults.

3·5. *Allowances for clothes.* In comparing the heights and weights of different series of children and adults, allowance has to be made for differences in the clothing worn by the subjects when measured. The allowances are required to adjust average values, and there can be no assurance that they are precisely correct when applied to a particular series. Relevant information is:

(*a*) With reference to a survey of male industrial workers: 'Height was taken without shoes, but with very thin heel-less slippers. Weights were taken with ordinary clothes on, 8 lb. being deducted from the gross weights registered, this value being the average of repeated weighings of male clothing' (Cathcart, Hughes & Chalmers, 1935).

(*b*) With reference to a survey of female industrial workers: 'Weight was taken without shoes, light heel-less slippers being provided.... Assessment was made of the weight of clothing worn by the average worker, and as a result 4 lb. was deducted from the gross weights obtained' (Cathcart, Bedale, Blair, Macleod & Weatherhead, 1927).

(*c*) The following means weights in kg. (1 kg. = 2·2 lb.) are given for the clothes of Alsatian boys measured in the spring (Schlesinger, 1917):

Age	7	8	9	10	11	12	13	14
Coat	0·2	0·2	0·25	0·3	0·4	0·5	0·6	0·6
'Trousers'	0·25	0·25	0·3	0·35	0·4	0·4	0·5	0·6
Underclothes	0·25	0·25	0·3	0·3	0·3	0·3	0·3	0·3
Boots	0·5	0·6	0·7	0·7	0·8	0·8	0·9	0·9
Total	1·2	1·3	1·55	1·65	1·9	2·0	2·3	2·4

(*d*) Measurements of R.A.F. aircrew clothed and unclothed (taken in December) give the following data:

Average excess of height in socks over height nude = 0·27 in. (520).
Average excess of weight clothed, without tunic or boots and with trouser pockets empty, over weight nude = 4·9 lb. (520).
Average weight of tunics = 2·5 lb.
Average weight of boots = 3·5 lb.

Measurements of the German 1946–7 survey are recorded for subjects in indoor clothing and without shoes. For various reasons some of the subjects were actually measured in other conditions, and the following corrections were applied to adjust such readings so that all records would be for subjects wearing the prescribed clothing (indoor clothing and without shoes):

Apparel	Correction for height (cm.)		Correction for weight (kg.)	
	Males	Females	Males	Females
Shoes	−2	−3	−1·0	−0·5
Overcoat and shoes	−2	−3	−3·5	−2·0
Heavy working dress	−2	—	−5·0	—
Religious sisters dress	—	−2	—	−3·0
Underwear only	—	—	+0·7	+0·7
None	—	—	+1·0	+1·0

It is not stated that these estimates were obtained (as average values) by actually measuring subjects wearing different sets of apparel, or by weighing garments. Most of them are in fairly close agreement with earlier estimates, but the additions to be made to weights when either underwear only or no clothing was worn are certainly too small.

In comparing different series mean heights and weights were adjusted when necessary by referring to the data given above (*a*)–(*d*) and choosing arbitrary corrections for clothing suitable to the circumstances. For children and adolescents the reduction is to measurements in indoor clothing and without shoes, so the German 1946–7 survey means are accepted as given. For weights of adults the reduction is to nude measurements, since most of the series give records for this condition, and the German means are adjusted accordingly. There is no assurance that allowances made for clothes are very close approximations to the unknown values in all cases, but errors on this account are not likely to be greater than about ¼ in. (6 mm.) for height and 1 lb. (0·45 kg.) for weight.

3·6. *Reduction of weights to constant heights.* In comparing the mean weights of two series it is necessary to make proper allowance for differences between their mean heights at different ages. This is best done for each age by using the regression coefficient of weight on height (giving increase in mean weight for unit increase in mean height), and reducing the mean weights of one of the series to values expected if its mean heights were those of the other series. Two sets of weights, one set estimated, for people having the same average heights are thus compared. Regression coefficients provided by the following British series were examined, all samples being for adequate numbers of individuals:

(*a*) male industrial workers, 1929–32, age groups 15 to over 60 years (Cathcart *et al.* 1935);

(*b*) female industrial workers, 1926, age groups 15 to over 50 years (Cathcart *et al.* 1927);

(*c*) Glasgow children divided into four groups representing different social grades, age groups 6–13 years (Elderton, 1914);

(*d*) East Sussex children, age groups 5–14 years (Dunstan, 1925);

(*e*) R.A.F. aircrew, 1944, age groups 20–40 years (unpublished).

The regression coefficients for these series are in good agreement, and there are clear differences between the values on account of sex, age and social class. In general they in-

crease with increase in mean weight. The age curves provided by the coefficients for each sex, and the relations between the curves for the two sexes, are very similar to typical age curves for weight (e.g. Fig. 8), female values only exceeding male for some adolescent ages. Series representing higher social classes tend to have greater coefficients for corresponding ages than series representing lower social classes, but differences of this kind are relatively small.

In practice it is not essential that regression coefficients used for the purpose considered should be derived from either of a pair of series compared, or that values used should be very close approximations to such ideal data. This is so because differences between the mean heights of the series at corresponding ages—the factor by which the regression

Table 1. *Data used in reducing weights to constant heights*

Age group (years): central values		6·5	7·5	8·5	9·5	10·5	11·5	12·5
Accepted regression coefficients, x (1 cm. $= x$ kg.)	Males	0·27	0·29	0·31	0·33	0·36	0·40	0·44
	Females	0·25	0·27	0·29	0·31	0·35	0·40	0·46
Heights (cm.) to which weights were reduced	Males	118	122	127	131	136	140	144
	Females	117	121	126	130	135	141	147

Age group (years): central values		13·5	14·5	15·5	16·5	17·5	18·5	All adult ages
Accepted regression coefficients, x (1 cm. $= x$ kg.)	Males	0·49	0·55	0·60	0·63	0·65	0·65	0·65
	Females	0·51	0·53	0·54	0·55	0·55	0·55	0·55
Heights (cm.) to which weights were reduced	Males	149	155	162	168	171	173	168·5
	Females	152	157	161	161	161	161	157·5

coefficients have to be multiplied—will nearly always be small. The reduced weights given in this paper were obtained by using a constant set of regression coefficients, these having been made up by balancing the evidence of the series listed above.

In comparisons with the German 1946–7 survey, the position is different for the immature series, on the one hand, and the adult on the other (see § 2·2). In the case of the former, absolute mean heights and weights are available for each year of age. It is convenient (at each age) to reduce the mean weights of all other series to the mean heights of the German series, these mean heights being given, together with the regression coefficients used, in

Table 1. In the case of the adults no absolute constants are available, but only mean weights reduced to the heights of 1685 mm. for men and 1575 mm. for women. The reduction is said to have been made for all age groups by applying the allowances (1 cm. in height) = (0·65 kg. in weight) for men, and (1 cm.) = (0·55 kg.) for women. These are presumably regression equations derived from correlation tables for height and weight, and they show good agreement with equations given by British series of adults. Accordingly, for all adult ages the German equations, and the heights to which the weights were reduced, were adopted. It may be noted that these heights (1685 mm. for men and 1575 mm. for women) are probably rather lower than the averages for German adults (cf. Fig. 7), so the reduced weights provided are probably rather less than the actual averages for the people. In comparisons where weights were adjusted the data given in Table 1 were used.

In view of the loss in precision—due chiefly to allowances which have to be made for clothes, and partly to the method of reduction to constant heights—no significance should be attributed to small differences between mean weights for different series. A difference greater than 2 lb., say, is almost certainly real, but a difference of 1 lb. may not be.

4. The heights and weights of Germans recorded in 1946–7

4·1. *Regional comparisons.* Comparisons can only be made for the months December 1946–March 1947 (children), or December 1946–February 1947 (adults). Pooled means were calculated for these periods, restricted to the years of age for which the data are most adequate, and for separate regions. The mean heights and weights for boys are plotted in Figs. 1 and 2, and the diagrams for girls give a similar impression of regional distinctions. Assuming that the order of variation was not exceptional, it can be estimated, by using standard deviations for other series, that most of the differences between the German means at corresponding ages are statistically insignificant, but some of the differences are almost certainly markedly significant. Taking weighted mean differences for ages between 6 and 16 years suggests the following order for both boys and girls:

North Rhine—Westphalia—Berlin—Hamburg—Hannover—Schleswig-Holstein.
(tallest) (shortest)

The weights as recorded give the similar orders:

Boys: North Rhine—Westphalia—Hamburg—Berlin—Hannover—Schleswig-Holstein.
(heaviest) (lightest)
Girls: North Rhine—Hamburg—Westphalia—Schleswig-Holstein—Hannover—Berlin.

The orders given by the mean weights are changed substantially when allowances are made for differences between the mean heights. The reduced weights, obtained by the method described in § 3·6, give:

Boys: Schleswig-Holstein—North Rhine—Westphalia—Hamburg—Hannover—Berlin.
(heaviest) (lightest)
Girls: North Rhine—Schleswig-Holstein—Hamburg—Westphalia—Hannover—Berlin.

For both boys and girls the differences between the extremes here are small (less than 1 kg.) for the younger ages and decidedly larger (several over 2 kg.) for ages over 12 years.

Fig. 2. Average weights (indoor clothing and without shoes) for regional series of German boys, December 1946–March 1947.

Fig. 1. Average heights for regional series of German boys, December 1946–March 1947.

Reduced mean weights for the adults are shown in Fig. 3 and they suggest the following sequences:

Men: Hannover, Hamburg and Schleswig-Holstein—Westphalia—North Rhine—Berlin.
(heaviest and no appreciable differences between levels) (lightest)

Women: Westphalia—North Rhine, Hamburg and Schleswig-Holstein—Hannover—Berlin.
(no appreciable differences)

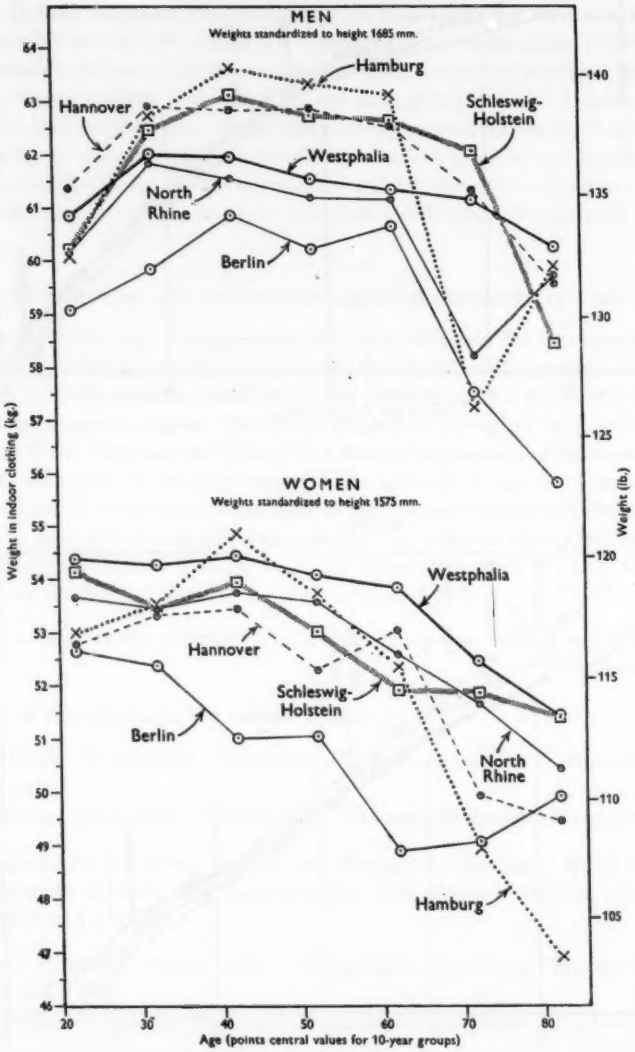

Fig. 3. Average weights (indoor clothing and without shoes) standardized for height of regional series of German men and women, December 1946–February 1947.

Excluding the Berlin series, the maximum differences between the reduced mean weights for the five other regions are of the order 2 kg. The only relation between the weights standardized for height which is consistent for both sexes and nearly all ages is that the people of Berlin were inferior to all the others. Regional differences between the other series are smaller in value and clearly of less significance.

4·2. *Secular comparisons: heights of children.* The total series considered here is that for the British zone of Germany excluding Berlin. The numbers of observations made in each month are roughly the same for the boys and girls and of the orders 10,000 for December 1946, 4000 for January 1947 and 11,000–13,000 for each of the months February–June 1947. A comparison of the mean heights for separate months failed to reveal any consistent secular trend. To obtain larger samples the seven months were divided into the two groups

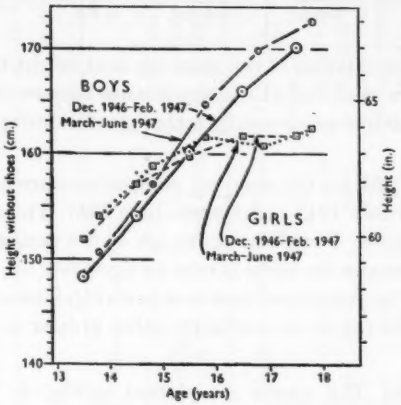

Fig. 4. Average heights of German boys and girls. (British zone except Berlin.)

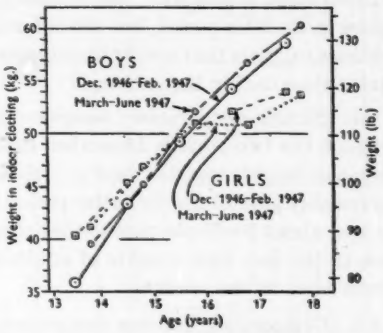

Fig. 5. Average weights (indoor clothing and without shoes) of German boys and girls. (British zone except Berlin.)

December 1946–February 1947 and March–June 1947, and the pooled means suggest the following conclusions:

Boys. Ignoring the two youngest ages for which the numbers are small, for ages under 13 years the two 'curves' are interlaced and the maximum divergence between them is about 15 mm. For ages over 13 (Fig. 4) the curves are separated and the divergence reaches a maximum of 25 mm. (1 in.) between ages 16 and 17 years, which is estimated to be markedly significant.

Girls. For ages under 13 the curves are interlaced and the maximum divergence is about 10 mm. For ages over 13 (Fig. 4) the means tend to be greater for the later period, but the maximum divergence is only 10 mm. and this may not be significant.

The general conclusion for both boys and girls is that there is no clear distinction between the growth rates for height of the population in the two periods for ages under 13; for ages over 13 there is indication, which is clear for boys and less clear for girls, that growth in height was faster in the later than in the earlier period. The distinction may be due partly to a normal seasonal fluctuation in the growth rate (see § 3·4).

4·3. *Secular comparisons: weights of children.* Comparisons of the age curves for the two periods leads to conclusions for weight similar to those for height. There are no clear distinctions for ages under 13. For ages over 13 (Fig. 5) the boys tend to be heavier for the

later period, the maximum divergence of the curves being about 1·4 kg. (3 lb.), but the girls tend to be practically the same or lighter for the later period. These comparisons relate to the actual weights recorded. Standardizing the means to the same height at each age, by the method described in § 3·6, has the effect of lessening the divergence of the weight curves and reversing the sign of the differences for most ages. The amounts in kg. by which the mean weights for the *earlier* period exceed those for the later after allowance has been made for differences in mean heights are:

Central value age group (years)	13·5	14·5	15·5	16·5	17·5
Boys	−2·4	+0·4	+0·8	+0·3	−0·1
Girls	+0·5	+0·5	+0·5	+0·9	+1·2

Interpreted in this way, the tendency was for children of the same age and height to be lighter in the later period, but the differences are small and of less significance because other evidence suggests that weight increments for children are normally rather greater during the winter than during the summer.

4·4. *Secular comparisons: weights of adults.* As for the children, comparisons are made between the two periods December 1946–February 1947 and March–June 1947. The data are mean weights standardized to constant heights. For males above age 40 the means are appreciably greater for the latter period: for females the same is true for ages over 70, but for ages about 30–70 the position is reversed. The distinction here is of more significance in view of the fact that weights of adults are believed to be normally rather greater in the winter than in the summer.

4·5. *Comparisons between occupational series.* The means are plotted in Fig. 6. The sequence: Very Heavy Workers (heaviest)—Heavy Workers—Miners—Moderately Heavy Workers and Normal Consumers (lightest) is clear for the men. For the women an order is less clear, but Normal Consumers tend to be appreciably lighter than Workers. The distinctions between the series must, of course, be influenced largely by processes of selection. For example, Very Heavy Workers (men) were as heavy for older as for younger age groups, which is not true for the total population (see Fig. 20). The distinction may be due to better feeding of the subgroup, or to the fact that middle-aged men below its average in physique and weight were liable to leave it because they were unable to continue heavy work, or both these factors may have been involved.

4·6. *Age curves for the total German series of children and adolescents.* (Table 3, Figs. 7, 8.) The series referred to are for the total British zone except Berlin. The age curves for height are almost coincident for ages under 11 years: for ages 11–15 the girls have the greater averages, this being the time when a clear puberty dip is shown by the curve for the boys. The curves provide evidence of the attainment of skeletal maturity at surprisingly young ages. The maximum average heights for males is seen to have been reached by 19·5 years, and for females by about 17 years. Extensive records for British men show that for the general population the maximum in question during the latter half of the nineteenth century was about 26 years, and that in the present century it tended to move to a younger age, being about 20 years to-day. The change is presumed to have been due to the better feeding of children. For some years before 1945 the German general population may have been fed better than the British, resulting in the attainment of maximum height, on the

Table 2. *Average weights taken at different periods of series of German men and women. Foods for British zone except Berlin.*

Nudet

Table 2. *Average weights taken at different periods of series of German men and women. Totals for British zone except Berlin.*

Years of birth	With indoor clothing and without footwear						Nude† (estimated)
	Months when measured						
	Dec. 1946–Feb. 1947		March–June 1947		Dec. 1946–June 1947		
	Mean (kg.)	c.v. age group	Mean (kg.)	c.v. age group	Mean (kg.)	c.v. age group	Mean (kg.)
	Men: weights standardized to height 1685 mm.						
1920–29	60·84 (13,382)*	22·0	61·00 (14,830)	22·3	60·92 (28,212)	22·2	57·4
1910–19	62·52 (11,357)	32·0	62·41 (12,472)	32·3	62·46 (23,829)	32·2	59·0
1900–09	62·44 (18,906)	42·0	62·89 (20,155)	42·3	62·67 (39,061)	42·2	59·2
1890–99	62·26 (11,908)	52·0	62·55 (12,992)	52·3	62·41 (24,900)	52·2	58·9
1880–89	62·07 (5,863)	62·0	62·33 (6,528)	62·3	62·21 (12,391)	62·2	58·7
1870–79	60·39 (2,140)	72·0	62·28 (2,544)	72·3	61·41 (4,684)	72·2	57·9
1860–69	59·68 (1,142)	82·0	61·12 (1,324)	82·3	60·45 (2,466)	82·2	56·9
1850–59	59·14 (71)	92·0	59·48 (52)	92·3	59·29 (123)	92·2	55·8
Totals	64,769	—	70,897	—	135,666	—	—
	Women: weights standardized to height 1575 mm.						
1920–29	53·67 (7,206)	22·0	53·41 (7,928)	22·3	53·53 (15,134)	22·2	51·5
1910–19	53·74 (9,763)	32·0	53·01 (10,899)	32·3	53·36 (20,662)	32·2	51·4
1900–09	53·50 (11,129)	42·0	53·30 (11,906)	42·3	53·40 (23,035)	42·2	51·4
1890–99	53·24 (8,239)	52·0	52·68 (8,564)	52·3	52·95 (16,803)	52·2	50·9
1880–89	52·60 (5,230)	62·0	51·91 (5,477)	62·3	52·25 (10,707)	62·2	50·2
1870–79	50·26 (3,738)	72·0	51·88 (4,317)	72·3	51·13 (8,055)	72·2	49·1
1860–69	49·80 (1,827)	82·0	51·27 (2,074)	82·3	50·58 (3,901)	82·2	48·6
1850–59	48·07 (208)	92·0	49·11 (202)	92·3	48·58 (410)	92·2	46·6
Totals	47,340	—	51,367	—	98,707	—	—

* A figure in brackets is the number of observations on which the mean is based and some repeated weights of the same people are included.
† Allowing for clothes 3·5 kg. (7·7 lb.) for men and 2 kg. (4·4 lb.) for women.

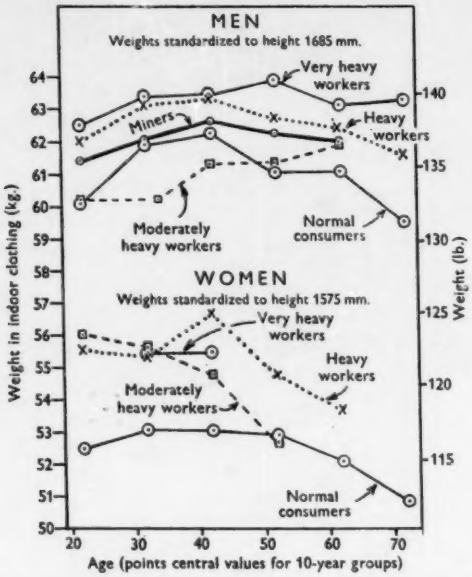

Fig. 6. Average weights (indoor clothing and without shoes) standardized for height of occupational series of German men and women. (British zone except Berlin, December 1946–June 1947.)

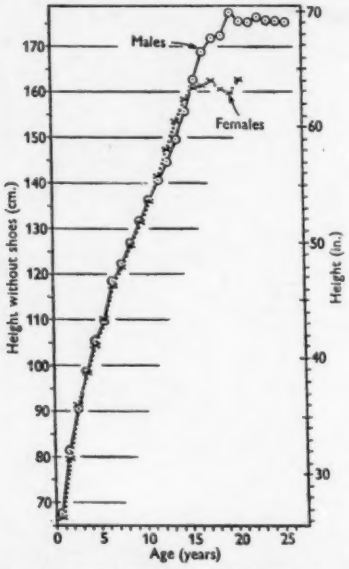

Fig. 7. Average heights of German males and females. (British zone except Berlin, December 1946–June 1947; data in Table 3.)

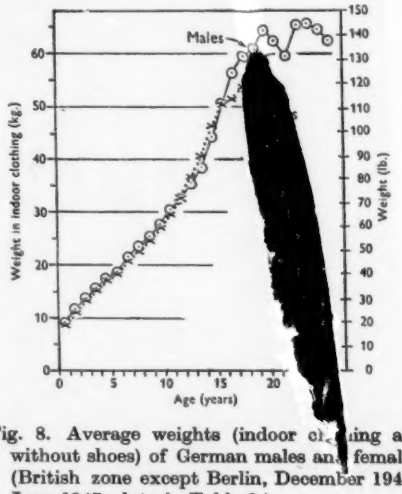

Fig. 8. Average weights (indoor clothing and without shoes) of German males and females. (British zone except Berlin, December 1946–June 1947; data in Table 3.)

Table 3. *Mean heights and weights of German boys and girls for the total British zone except Berlin*

A. *Measurements as recorded in indoor clothing and without shoes*

Age (years): central values	No. of measurements*		Mean height (mm.)		Mean weight (kg.)	
	Boys	Girls	Boys	Girls	Boys	Girls
0·7	153	135	676·7	668·3	9·03	8·49
1·7	116	71	814·9	798·6	11·77	11·11
2·7	648	448	907·7	914·7	13·94	13·26
3·7	1,055	1,076	985·5	984·1	15·84	15·34
4·7	1,451	1,407	1051·7	1040·7	17·54	17·12
5·7	1,673	1,444	1099·1	1101·2	18·87	18·29
6·7	3,714	3,112	1184·9	1174·6	21·67	20·79
7·7	8,868	8,143	1221·9	1219·5	23·23	22·39
8·7	8,390	8,569	1269·8	1264·9	25·35	24·55
9·7	9,523	9,420	1317·5	1315·0	27·83	27·01
10·7	10,879	11,414	1362·6	1360·2	30·28	29·74
11·7	10,246	11,048	1408·2	1413·7	32·81	32·83
12·7	8,752	9,057	1447·5	1472·1	35·71	36·66
13·7	5,167	4,896	1498·1	1534·1	38·28	40·92
14·7	3,277	3,076	1558·5	1581·0	44·48	46·23
15·7	1,750	1,436	1627·5	1609·2	50·95	50·48
16·7	834	547	1689·1	1610·6	56·20	51·32
17·7	452	416	1719·3	1621·5	59·87	53·81
18·7	119	216	1720·3	1606·2	61·25	54·04
19·7	74	57	1774·4	1598·2	64·56	53·24
20·7	33	32	1755·8	1622·7	62·40	52·18
21·7	19	—	1750·1	—	59·47	—
22·7	43	—	1763·0	—	65·27	—
23·7	37	—	1755·8	—	65·75	—
24·7	44	—	1753·9	—	64·58	—
25·7	43	—	1750·2	—	62·59	—
Totals	77,360	76,020	—	—	—	—

* These are the numbers of observations on which the means are based and some repeated measurements of the same individuals are included.

B. *Estimated nude weights reduced to height* 1685 *mm. for males and* 1575 *mm. for females*

Age (years): central values	Males†	Females‡
16·7	52·93 (834)	47·61 (547)
17·7	54·39 (452)	49·25 (416)
18·7	55·46 (119)	50·32 (216)
19·7	56·55 (74)	49·96 (57)

† Allowances for clothes 3 kg. (6·6 lb.) for age group 16·7, 3·25 for 17·7 and 3·5 for groups 18·7 and 19·7.

‡ Allowances for clothes 1·75 kg. (3·9 lb.) for age group 16·7, 2 for 17·7, 18·7 and 19·7.

average, at a still younger age. Restrictions thereafter would not be expected to modify the form of the age curve appreciably until after a lapse of some years.

The weight curves cross at about ages 11·5 and 15·5 years. Compared with pre-war data, both are peculiar in showing a marked decline in weight increments after age 18.

5. Comparisons of the post-war and earlier German series

5·1. Records of the heights and weights of German children living in the British zone of Germany are fairly abundant for years before 1939 and a few of the series were selected for comparison here. They are for:

(*a*) Boys and girls measured in Berlin about 1900, data for 'higher schools' (*Gymnasien*) and 'lower schools' (*Gemeindeschulen*) being given separately (Rietz, 1903).

(*b*) Boys and girls in a Berlin orphanage measured in 1919 (Davidsohn, 1919).

(*c*) Boys and girls (heights only) measured in 'lower schools' (*Volkschulen*) in Kiel and other towns in Holstein in 1902 (Ranke, 1905).

(*d*) Boys only measured in 'higher schools' in Hamburg about 1878 (Kotelmann, 1879, quoted by Ranke, 1905).

Means for these series are given in Tables 4 and 5 and those for the Berlin children are plotted, together with the means for the 1946–7 series, in Figs. 9–12.

5·2. As expected, the age curves for the pre-war Berlin series make clear distinctions between the social classes represented. It is surprising, however, to find that the curves for the general population in 1946–7 indicate a standard between those of the earlier 'higher school' children, on the one hand, and of the 'lower school' and orphanage children, on the other. This is so for both heights and weights of both boys and girls. The present-day Schleswig-Holstein series is also superior to earlier 'lower school' children of Holstein in the case of heights of boys and girls; and for Hamburg series the 1946–7 means exceed those of 'higher school' boys in 1878. The standards of the post-war German children are evidently not markedly depressed.

5·3. The average weights are best compared by reducing them at each age to a constant height (see § 3·6). The heights used for this purpose are those of the total 1946–7 series for the British zone except Berlin, and the weights of all the pre-war German series reduced to these heights are given in Table 7 and plotted in Figs. 17 and 18. In general the post-war series show standards of weights, after allowance is made for differences in height, which do not compare unfavourably with those of pre-war lower social grades of the German population (see also § 6·3).

5·4. The writer has been unable to find any records of large pre-war series of German adults suitable for comparison with the 1946–7 data. Maps compiled by anthropologists (Coon, 1939, and earlier authorities) show average heights for the male adult populations of different parts of the British zone of Germany ranging from 168 to 173 cm. The mean for the region in 1946–7 is about 176 cm. (Fig. 7). A probable explanation of the superiority of the later estimate is that it represents the maximum of the age curve (reached to-day about 19½ years), whereas the earlier estimates were derived mainly from data for conscripts (ages about 18–20) at times when the maximum of the age curve was normally not attained until some age near 25 years. It is known for the general British population that the age in question changed from about 26 years in the latter half of the nineteenth century to about 20 years to-day, while the maximum mean height attained remained unchanged in the past hundred years.

Table 4. *Average heights (without shoes) in mm. for pre-war series of German children*

Age (years): central values	Boys					Girls				
	Berlin			Holstein	Hamburg	Berlin				Holstein
	'Higher schools' c. 1900	'Lower schools' c. 1900	Orphanage 1919	'Lower schools' 1902	'Higher schools' c. 1878	'Higher schools' c. 1900	'Lower schools' c. 1900	Orphanage 1919		'Lower schools' 1902
6	—	—	988 (86)	—	—	—	—	1006 (74)		—
6·5	1183 (45)	1136 (128)	—	1059 (47)	—	1190 (14)	1119 (110)	—		1073 (43)
7	—	—	1068 (82)	—	—	—	—	1107 (45)		—
7·5	1220 (101)	1172 (189)	—	1117 (185)	—	1227 (42)	1173 (159)	—		1136 (81)
8	—	—	1142 (58)	—	—	—	—	1135 (55)		—
8·5	1273 (156)	1214 (198)	—	1169 (189)	—	1272 (37)	1217 (164)	—		1165 (108)
9	—	—	1175 (74)	—	—	—	—	1185 (47)		—
9·5	1312 (168)	1265 (192)	—	1215 (164)	1286	1310 (54)	1250 (182)	—		1223 (114)
10	—	—	1238 (57)	—	—	—	—	1218 (45)		—
10·5	1357 (181)	1309 (198)	—	1272 (178)	1307	1357 (71)	1306 (182)	—		1273 (110)
11	—	—	1279 (48)	—	—	—	—	1268 (46)		—
11·5	1395 (209)	1353 (211)	—	1308 (156)	1351	1412 (69)	1357 (185)	—		1318 (125)
12	—	—	1327 (56)	—	—	—	—	1303 (33)		—
12·5	1451 (189)	1397 (181)	—	1354 (131)	1399	1478 (65)	1408 (169)	—		1369 (127)
13	—	—	1352 (48)	—	—	—	—	1368 (29)		—
13·5	1506 (143)	1447 (162)	—	1397 (131)	1431	1521 (73)	1481 (180)	—		1411 (126)
14	—	—	1404 (53)	—	—	—	—	1409 (32)		—
14·5	1560 (158)	1466 (37)	—	1453 (117)	1488	1566 (62)	1505 (34)	—		1462 (105)
15·5	1624 (140)	—	—	1496 (76)	1542	1580 (46)	—	—		1475 (28)
16·5	1658 (117)	—	—	—	1616	—	—	—		—
17·5	1690 (70)	—	—	—	1669	—	—	—		—
18·5	1710 (40)	—	—	—	1684	—	—	—		—
19·5	1711 (23)	—	—	—	1668	—	—	—		—
Totals	1,740	1,496	562	1,374	*c.* 500	533	1,365	406		967

Table 5. *Average weights (in indoor clothing and without shoes) in kg. for pre-war series of German children*

Age (years): central values	Boys				Girls		
	Berlin			Hamburg	Berlin		
	'Higher schools' *c.* 1900	'Lower schools' *c.* 1900	Orphanage* 1919	'Higher schools' *c.* 1878	'Higher schools' *c.* 1900	'Lower schools' *c.* 1900	Orphanage* 1919
6	—	—	16·7 (86)	—	—	—	16·4 (74)
6·5	22·3 (45)	20·1 (128)	—	—	22·5 (14)	19·6 (110)	—
7	—	—	20·6 (82)	—	—	—	19·3 (45)
7·5	23·7 (101)	21·6 (189)	—	—	24·3 (42)	21·6 (159)	—
8	—	—	21·1 (58)	—	—	—	19·8 (55)
8·5	26·2 (156)	23·3 (198)	—	—	26·1 (37)	23·3 (164)	—
9	—	—	22·3 (74)	—	—	—	22·3 (47)
9·5	27·8 (168)	25·7 (192)	—	26·9	27·8 (54)	24·7 (182)	—
10	—	—	25·5 (57)	—	—	—	24·0 (45)
10·5	30·6 (181)	27·6 (198)	—	28·3	32·1 (71)	27·5 (182)	—
11	—	—	28·3 (48)	—	—	—	26·5 (46)
11·5	33·1 (209)	30·0 (211)	—	30·7	34·4 (69)	30·3 (185)	—
12	—	—	30·4 (56)	—	—	—	28·2 (33)
12·5	37·1 (189)	32·9 (181)	—	33·9	40·5 (65)	34·4 (169)	—
13	—	—	32·9 (48)	—	—	—	33·2 (29)
13·5	41·6 (143)	36·5 (162)	—	35·8	43·1 (73)	39·3 (180)	—
14	—	—	34·7 (53)	—	—	—	37·5 (32)
14·5	46·1 (158)	37·5 (37)	—	41·0	49·7 (62)	43·1 (34)	—
15·5	51·7 (140)	—	—	45·9	51·2 (46)	—	—
16·5	56·3 (117)	—	—	51·9	—	—	—
17·5	59·1 (70)	—	—	56·9	—	—	—
18·5	64·4 (40)	—	—	60·4	—	—	—
19·5	65·5 (23)	—	—	61·8	—	—	—
Totals	1,740	1,496	562	*c.* 500	533	1,365	406

* Nude weight given: allowances made for clothes as in table in § 3·5*c*.

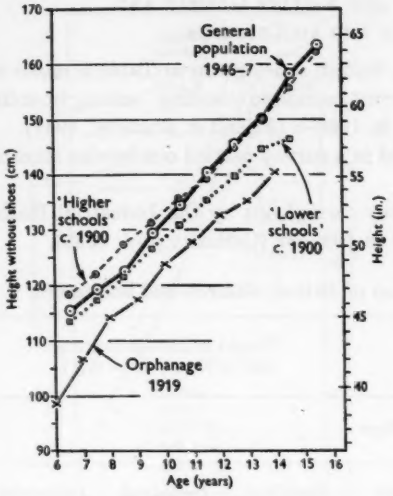

Fig. 9. Average heights for series of boys measured in Berlin. (Data for 1946–7 series in Table 3, and for other series in Table 4.)

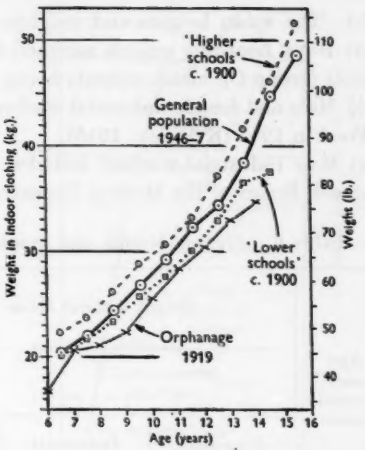

Fig. 10. Average weights (indoor clothing and without shoes) for series of boys measured in Berlin. (Data for 1946–7 series in Table 3, and for other series in Table 5.)

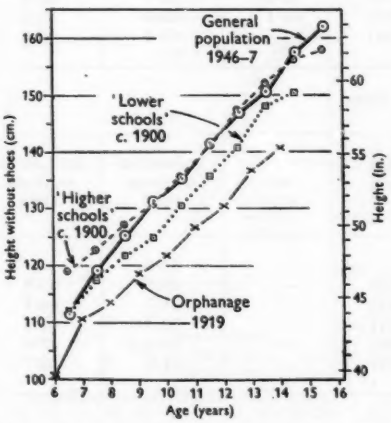

Fig. 11. Average heights for series of girls measured in Berlin. (Data for 1946–7 series in Table 3, and for other series in Table 4.)

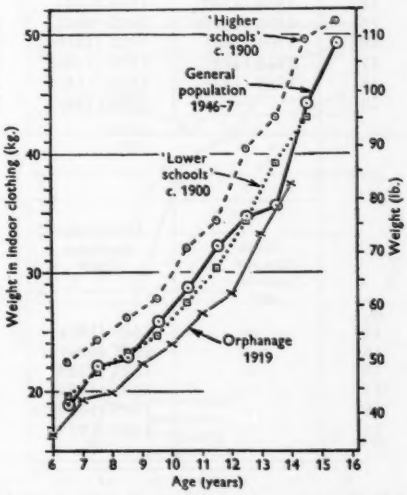

Fig. 12. Average weights (indoor clothing and without shoes) for series of girls measured in Berlin. (Data for 1946–7 series in Table 3, and for other series in Table 5.)

* Nude weight given: allowances made for clothes as in table in § 3·5c.

6. Comparisons of the post-war and earlier German and British series of children and adolescents

6·1. The mean heights and weights of recent British series given in Table 6 relate to:

(*a*) Boys' boarding schools assigned to the 'lowest economic standing' among boarding schools (group C), measurements being recorded in 1936–8 (Friend & Bransby, 1947).

(*b*) Male and female industrial workers included in a survey carried out by the Ministry of Food in 1943 (Kemsley, 1945).

(*c*) Male industrial workers included in a survey carried out by the Industrial Health Research Board of the Medical Research Council in 1929–32 (Cathcart *et al.* 1935).

Table 6. *Average heights and weights for series of British children and adolescents*

Age (years): central values	Height without shoes (mm.)			Weight in indoor clothing and without shoes (kg.)		
	Boys					
	Boarding schools 1936–8	Industrial workers 1943	Industrial workers 1929–32	Boarding schools 1936–8	Industrial workers 1943	Industrial workers 1929–32
11	1427 (284)	—	—	35·0	—	—
12	1443 (563)	—	—	36·8	—	—
13	1504 (888)	—	—	40·9	—	—
14	1565 (1537)	1558 (222)	—	45·9	45·9	—
15	1628 (1573)	1580 (806)	1540 (139)	51·7	49·1	44·9
16	1704 (544)	1637 (1310)	1603 (206)	58·9	54·0	50·7
17	1745 (412)	1673 (1560)	1656 (295)	63·6	57·7	55·8
18	1750 (152)	1684 (1240)	1686 (329)	65·1	59·9	59·2
19	—	1693 (729)	1691 (335)	—	61·1	60·6
	Girls					
	—	Industrial workers 1943	Industrial workers 1926	—	Industrial workers 1943	Industrial workers 1926
14	—	1545 (189)	—	—	46·5	—
15	—	1557 (698)	1557 (152)	—	49·5	45·8
16	—	1575 (1332)	1566 (213)	—	51·7	48·1
17	—	1579 (1622)	1570 (257)	—	52·8	49·5
18	—	1579 (1620)	1582 (259)	—	53·5	52·0
19	—	1582 (1377)	1585 (269)	—	54·0	52·7

(*d*) Female industrial workers included in a survey carried out by the same Board in 1926 (Cathcart *et al.* 1927).

6·2. The age 'curves' for these British and for the 1946–7 German (British zone except Berlin) series are shown in Figs. 13–16. For heights and weights of boys the German curves fall between those for the British pre-war boarding schools and industrial series, and the

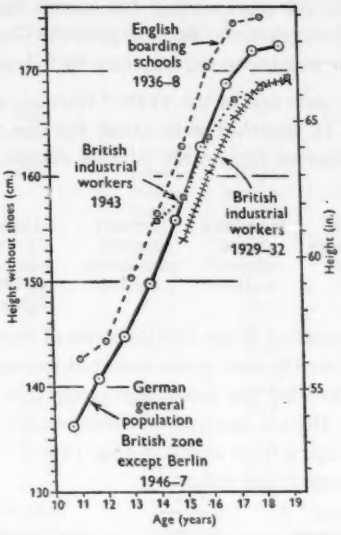
English boarding schools 1936–8
British industrial workers 1943
British industrial workers 1929–32
German general population British zone except Berlin 1946–7
Height without shoes (cm.)
Height (in.)
Age (years)

Fig. 13. Average heights for three British and a German series of boys. (Data for British series in Table 6, and for the German in Table 3.)

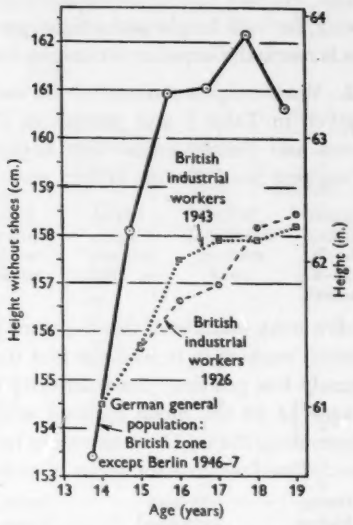
British industrial workers 1943
British industrial workers 1926
German general population: British zone except Berlin 1946–7
Height without shoes (cm.)
Height (in.)
Age (years)

Fig. 14. Average heights for two British and a German series of girls. (Data for British series in Table 6, and for the German in Table 3.)

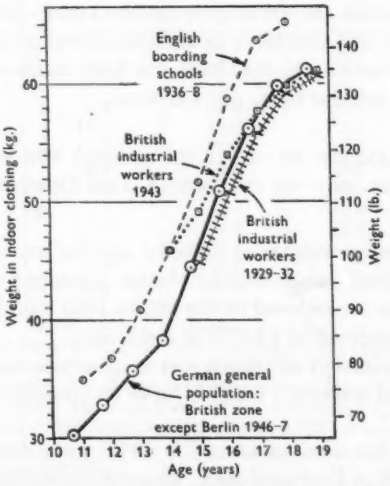
English boarding schools 1936–8
British industrial workers 1943
British industrial workers 1929–32
German general population: British zone except Berlin 1946–7
Weight in indoor clothing (kg.)
Weight (lb.)
Age (years)

Fig. 15. Average weights (indoor clothing and without shoes) for three British and a German series of boys. (Data for British series in Table 6, and for the German in Table 3.)

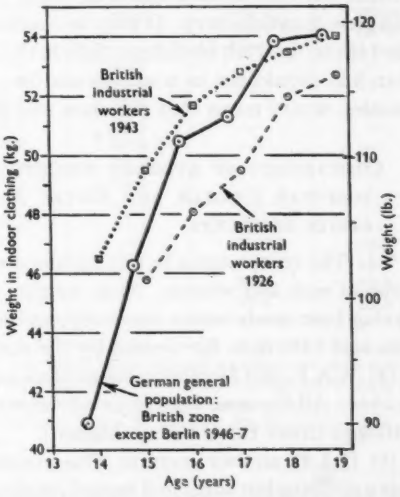
British industrial workers 1943
British industrial workers 1926
German general population: British zone except Berlin 1946–7
Weight in indoor clothing (kg.)
Weight (lb.)
Age (years)

Fig. 16. Average weights (indoor clothing and without shoes) for two British and a German series of girls. (Data for British series in Table 6, and for the German in Table 3.)

three have very similar forms. It may be noted that the 1943 British series of male industrial workers differs appreciably from the 1929–32 representing the same class in showing, for both height and weight, greater means for ages under 18. For girls the German series is markedly superior for height, but inferior for weight, to the wartime British series.

6·3. Mean weights reduced to the same height at each age of the 1946–7 German series are given in Table 7 and plotted in Figs. 17 and 18 together with those for the other German and British series. For boys the various curves run nearly parallel courses and they suggest the following order:

English boarding schools 1936–8 (heaviest)	British industrial workers 1943	Berlin 'higher schools' *c.* 1900	British industrial workers 1929–32	Berlin orphanage 1919	Hamburg 'higher schools' *c.* 1878	German general population 1946–7	Berlin 'lower schools' *c.* 1900 (lightest)

Noteworthy features of this sequence are the high position of the British series of wartime industrial workers in it, and the fact that the post-war German series does not occupy an extremely low position. Its inferiority is most marked for the oldest age group (age 18). For ages 14–18 the mean reduced weights for the British wartime workers exceed the corresponding 1946–7 German values by amounts ranging from about 2–3 kg. (4·5–7·7 lb.).

The reduced weights for series of girls (Fig. 18) suggest the order:

Berlin 'higher schools' *c.* 1900 (heaviest)	British industrial workers 1943	Berlin 'lower schools' *c.* 1900	Berlin orphanage 1919	German general population 1946–7	British industrial workers 1926 (lightest)

A British pre-war standard is here below the post-war German, but the level of the British wartime series is decidedly high. The latter conclusion, which applies to both boys and girls, is satisfactory. It may be concluded that for the age range considered here—from the 14th to the 19th birthday—falls in the British 1943 standards for industrial workers less than 3 lb. would not be a serious matter. Reductions up to that limit, for both males and females, would mean that the class was still not inferior to its pre-war level.

7. Comparisons of average weights standardized to the same heights for the post-war German and Royal Air Force and other British and Dominion series of adults

7·1. The comparisons in this section are of average weights at different ages for various series of men and women. Nude weights are treated, allowances for clothes given in § 3·5 having been made where necessary, and all means are reduced to the height 1685 mm. for men and 1575 mm. for women by the method described in § 3·6. The series are:

(*a*) R.A.F. and Dominion series of aircrew. The survey of heights and weights was made in 1944. All the men were in an advanced stage of training (at O.T.U.'s) or on operational duties in Great Britain (unpublished).

(*b*) R.A.F. aircrew recruits. The majority of the men, measured in 1942, were direct entry civilians but some had served previously as R.A.F. ground staff. Many of the subjects of this survey must have been remeasured in the 1944 survey (*a*) above, though the exact proportion is unknown (Morant, 1943).

(*c*) R.A.F. aircrew in training. When measured in 1944 the men had just returned from training overseas (Canada and South Africa), (Morant & Gilson, 1945).

Table 7. *Average weights (in indoor clothing and without shoes) in kg. reduced at each age to the average height of the German (British zone except Berlin) 1946–7 series for various other German and British series of children and adolescents**

Age (years): central values	Boys							
	German					British		
	General population 1946–7	Berlin 'Higher schools' c. 1900	Berlin 'Lower schools' c. 1900	Berlin orphanage 1919	Hamburg 'Higher schools' c. 1878	Boarding schools 1936–8	Industrial workers 1943	Industrial workers 1929–32
6·5	21·2	22·2	21·3	22·7	—	—	—	—
7·5	23·1	23·7	23·0	24·2	—	—	—	—
8·5	25·4	26·1	25·0	25·2	—	—	—	—
9·5	27·5	27·7	27·2	27·4	27·7	—	—	—
10·5	30·0	30·7	29·4	30·6	30·2	—	—	—
11·5	32·2	33·3	31·9	33·2	32·7	34·5	—	—
12·5	35·0	36·6	34·8	36·1	35·7	37·4	—	—
13·5	38·0	40·8	38·6	39·3	38·7	41·2	—	—
14·5	43·4	45·5	42·1	42·2	44·4	46·2	46·5	—
15·5	49·9	51·5	—	—	50·6	52·5	52·2	50·7
16·5	55·4	57·7	—	—	55·9	58·4	57·7	56·4
17·5	59·1	60·4	—	—	59·6	62·4	60·9	60·0
18·5	61·1	65·7	—	—	63·4	—	63·2	62·6

Age (years): central values	Girls					
	German				British	
	General population 1946–7	Berlin 'Higher schools' c. 1900	Berlin 'Lower schools' c. 1900	Berlin orphanage 1919	Industrial workers 1943	Industrial workers 1926
6·5	20·3	22·0	20·9	20·7	—	—
7·5	22·2	23·8	22·6	21·9	—	—
8·5	24·2	25·8	24·5	23·9	—	—
9·5	26·4	27·5	26·3	26·2	—	—
10·5	29·0	31·9	29·0	29·0	—	—
11·5	32·0	34·3	32·4	32·3	—	—
12·5	35·9	40·1	37·3	36·9	—	—
13·5	40·0	43·0	41·3	42·1	—	—
14·5	45·4	49·9	46·5	47·1	49·0	—
15·5	49·9	52·8	—	—	53·0	49·5
16·5	51·3	—	—	—	54·1	51·1
17·5	53·5	—	—	—	53·9	52·7
18·5	54·0	—	—	—	55·4	53·8

* Another series of interest is of Belgian children of a working-class district of Brussels measured in the last three months of 1942, 1943 and 1944 (Ellis, 1945). To obtain large enough numbers at each age the data for the three years were pooled, giving n's ranging from 45 to 309, most representing more than 200 individuals. It is shown in the paper describing the records that the wartime children were *above* pre-war Belgian standards for height, and of the same order as the pre-war standards for weight. Reducing the mean weights (kg.) to the heights of the German 1946–7 series gives:

Age (years): central values	6·5	7·5	8·5	9·5	10·5	11·5	12·5	13·5	14·5	15·5
Boys	21·1	22·7	24·9	27·3	30·3	32·3	35·4	39·0	44·1	50·3
Girls	20·4	22·0	24·4	26·3	29·1	32·7	36·9	42·4	48·0	52·2

Compared with the series in the table above these reduced weights (in indoor clothing and without shoes) are extremely low for boys aged 6, 7 and 8, and low but not extreme in nearly all other cases

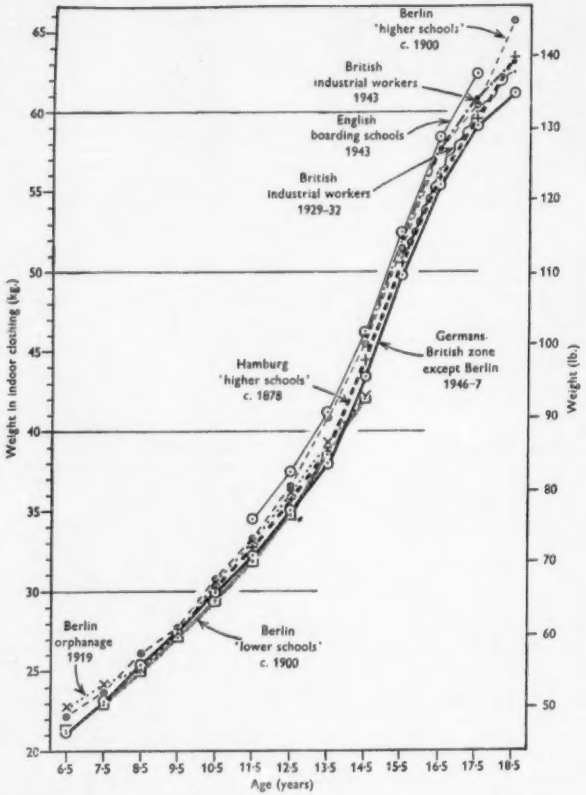
Berlin 'higher schools' c. 1900
British industrial workers 1943
English boarding schools 1943
British industrial workers 1929–32
Germans British zone except Berlin 1946–7
Hamburg 'higher schools' c. 1878
Berlin orphanage 1919
Berlin 'lower schools' c. 1900
Weight in indoor clothing (kg.)
Weight (lb.)
Age (years)

Fig. 17. Average weights (indoor clothing and without shoes) of German and British series of boys reduced to the average heights of the German 1946–7 series. (Data in Table 7.)

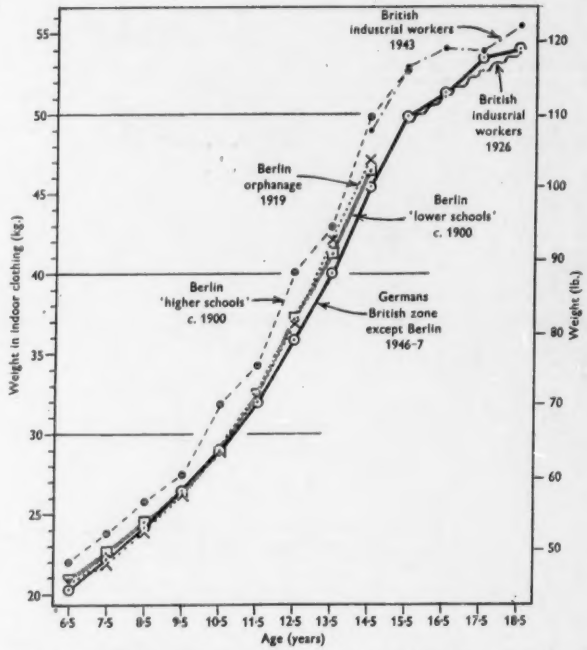
British industrial workers 1943
British industrial workers 1926
Berlin orphanage 1919
Berlin 'lower schools' c. 1900
Berlin 'higher schools' c. 1900
Germans British zone except Berlin 1946–7
Weight in indoor clothing (kg.)
Weight (lb.)
Age (years)

Fig. 18. Average weights (indoor clothing and without shoes) of German and British series of girls reduced to the average heights of the German 1946–7 series. (Data in Table 7.)

(*d*) R.N. pilots. The measurements used were those of the men when they were recruited for the Royal Navy from 1940 to 1946 (Morant, 1947).

(*e*) R.A.F. ground staff. The series represents a random sample of men recruited from 1940 to 1944, the measurements being those recorded at their first medical examinations (unpublished).

(*f*) Conscripts examined from 1 November 1917 to 31 October 1918. The data are for men examined by recruiting Medical Boards of the West Midland Region of England (Ministry of National Service, 1920). Those who were rejected for service in the Forces are included. At the time many who had previously been rejected were recalled for examination.

(*g*) Series of pre-war industrial workers. The surveys, carried out for the Industrial (Fatigue) Health Research Board of the Medical Research Council, were of women in 1926 (Cathcart *et al.* 1927) and of men in 1929–32 (Cathcart *et al.* 1935).

(*h*) Series of wartime industrial workers. The survey was carried out for the Ministry of Food in 1943 (Kemsley, 1945). Mean heights and weights given in the report were used to obtain the reduced weights shown in Figs. 20 and 21.

It should be noted that the Service series (*b*), (*d*), (*e*) and (*f*) are made up by men of whom all, or most, were civilians up to the time measurements were recorded.

7·2. The reduced weights for all the British and for the post-war German series are plotted in Figs. 19 and 20 (men) and 21 (women). In the case of the men there are far more series for ages up to 40 than for ages over 40. Considering the younger age range (17–40 years) only the series suggest the following order, though the sequence is not exactly the same for all ages within the range:

Operational aircrew: New Zealand	heaviest
Australian	
Canadian	
R.A.F.: Aircrew in training	
Operational aircrew	
R.N. pilot recruits (most civilian)	
R.A.F. aircrew recruits (most civilian)	
R.A.F. ground-staff recruits	
Industrial workers, 1943	
Employed industrial workers, 1929–32	
Germans, 1946–7	
Conscripts, 1917–18 and unemployed industrial workers, 1929–32	lightest

This sequence is clearly significant. During the war, aircrew were the best-fed section of the British population, though the R.A.F. did not reach the high weight levels of the Dominion aircrew. The superiority of the New Zealand series is suggestive in view of the high nutritional and health standards of that country. The slight superiority of the aircrew in training over the British operational aircrew might be due to the fact that the former were measured immediately after their return from overseas; or the distinction might be partly due to the loss in weight, on the average, during the earlier stages of wartime operational tours, which was found to be of the order 1·5 lb. (Reid, 1947). Aircrew recruits (R.N. pilots and R.A.F.) come next, followed by R.A.F. ground-staff recruits. When plotted together, the R.A.F. ground-staff recruits (Fig. 19) are seen to have greater mean

German and British series of girls reduced to the average heights of the German 1946–7 series. (Data in Table 7.)

German and British series of boys reduced to the average heights of the German 1946–7 series. (Data in Table 7.)

weights than the wartime and pre-war industrial workers (Fig. 20) for ages up to about 28 years, but smaller means for all later ages represented. The younger men accepted for service, chiefly on medical grounds, must have been heavier, on the average, than the

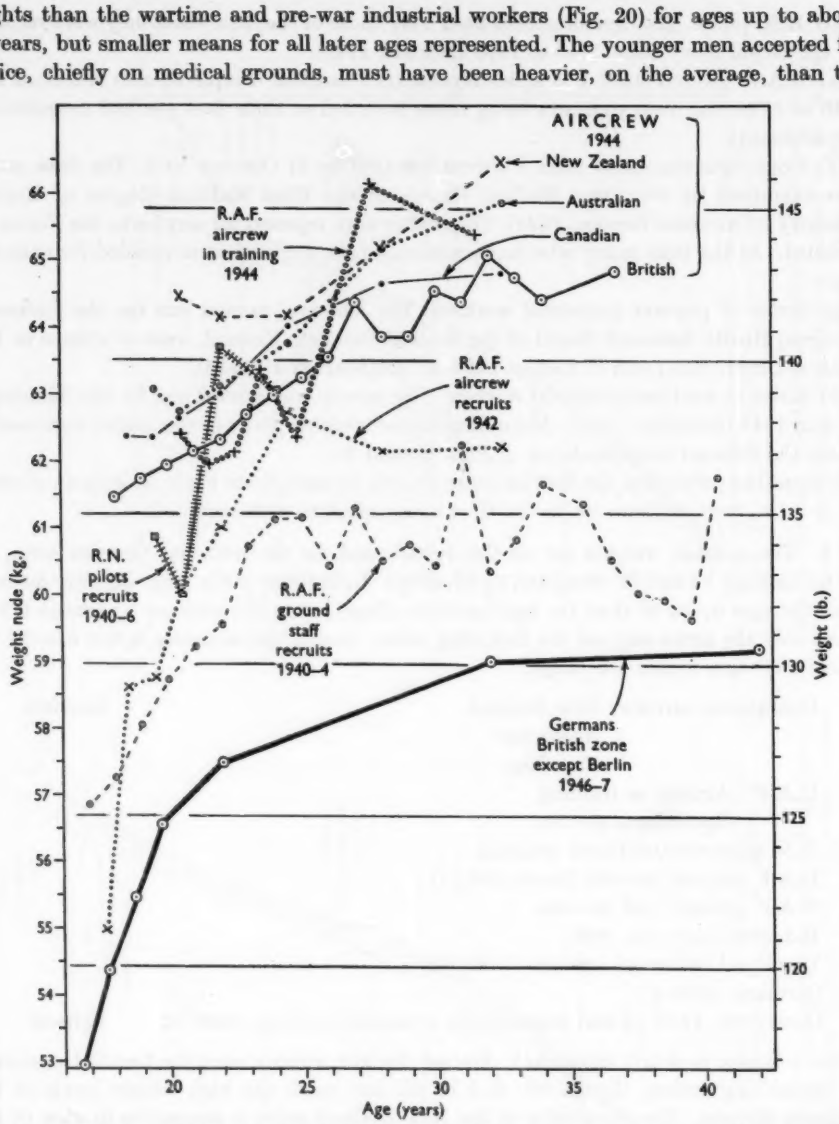

Fig. 19. Average weights (nude) standardized to height 1685 mm. for the post-war German series of civilians and series of R.N. and R.A.F. personnel. The smallest series are of R.N. pilots ($n = 200$), R.A.F. aircrew in training (529) and New Zealand aircrew (550), all the others being of 1600 or more men.

population from which they were drawn, but for the older men the position was reversed. For men up to age 40 it is surprising to find that the post-war German series has to be classed rather above two earlier British ones. Compared with them it shows superiority in average weights for ages up to about 25 years, and the same level as theirs for ages 25–40.

For ages over 40 (Fig. 20) the situation is markedly different. The German series clearly falls to the lowest place and the wartime industrial workers fall below the pre-war employed workers of the same class, though remaining above the low levels of the pre-war unemployed

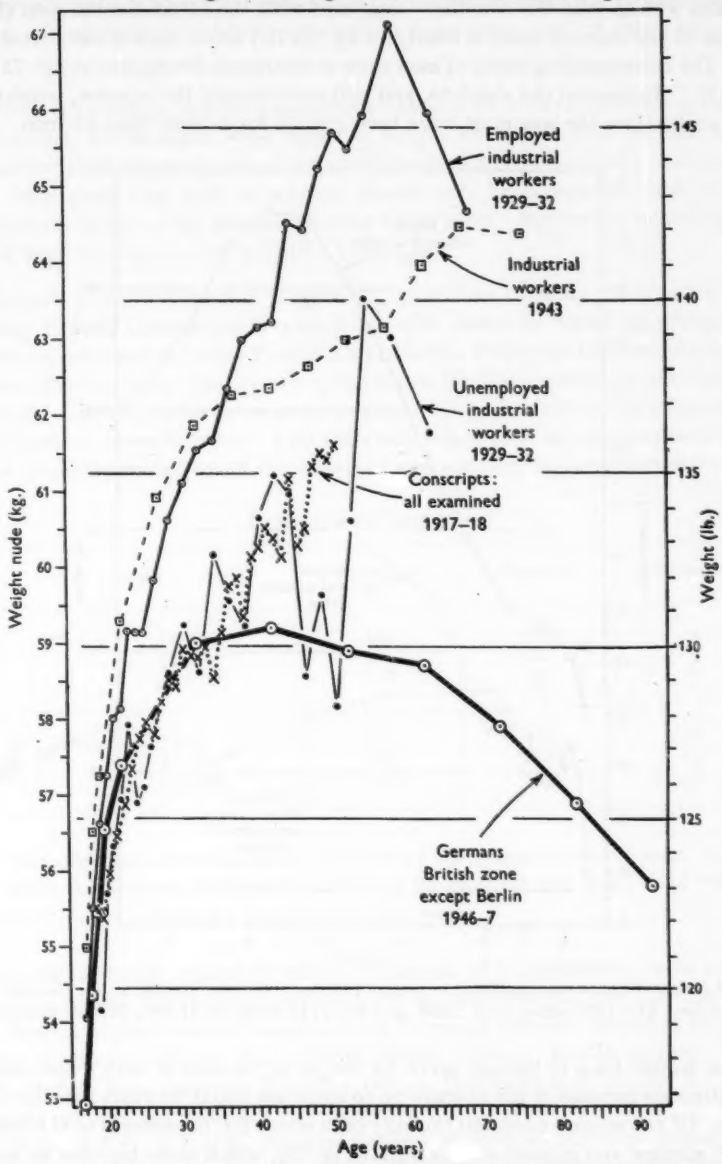

Fig. 20. Average weights (nude) standardized to height 1685 mm. for the post-war German and British series of civilian men. The smallest series is of 1,328 unemployed industrial workers and each of the other four relates to more than 10,000 men.

industrial workers and the 1917–18 conscripts. Food rationing affects most markedly the weights of the middle aged and aged.

This is shown again by the few female series (Fig. 21). The maximum of the age curve for the German women is in the twenties—compared with the forties for the men (Fig. 20)—and at age 57 the German mean is nearly 10 kg. (22 lb.) below that of the British wartime workers. The corresponding series of men show a maximum divergence at age 75 of nearly 7 kg. (15 lb.). In general the absolute, and still more clearly the relative, weight losses of German adults since the war must have been greater for women than for men.

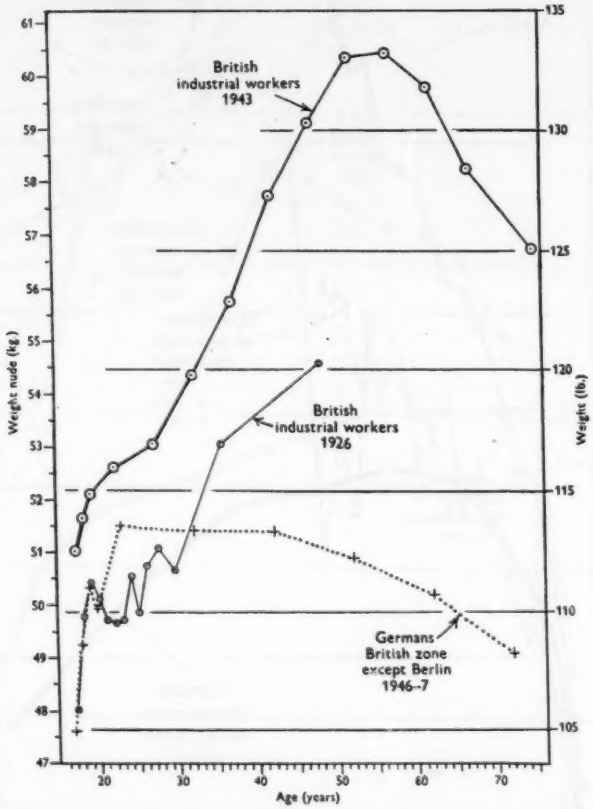

Fig. 21. Average weights standardized to height 1575 mm. for two British and a German series of women. The 1926 series is of 3,000, and the 1943 series of 31,200, British women.

7·3. The normal form of the age curve for weight in the case of both males and females shows continuous increase of the average up to some age about 60 years and then a decline in old age. Of the series considered the only ones which fail to conform to this pattern are the R.A.F. aircrew and ground-staff recruits (Fig. 19), which show increase to age 25 and then remain at about the same level, the German series of men (Fig. 20, maximum about 40 years) and women (Fig. 21, decline from ages 19–20 and maximum in early twenties),

and the pre-war British series of women industrial workers (Fig. 21, abnormal decline between ages 19 and 24). All except the last of these series had been subject to civilian rationing, but the 1943 men and women industrial workers, who show the normal form of curve, were also subject to it. The peculiarities of the recent series are probably due to the interaction of rationing, of medical selection in some cases, and of other imponderable factors.

7·4. The data presented provide standards which may be of use in future comparisons. It might be asked, for example, what losses in weight for a particular civilian or Service population for which repeated records are available should be considered a matter of serious concern. It is clear that such a question should only be considered with reference to particular years of age or age groups covering a few years. Statements regarding all men, say, or all boys, would be of no scientific value.

7·5. Records of average weights appreciably lower than those for the post-war Germans are not easy to find. Comparison is made below with means for broad age groups given for United States prisoners in Santo Tomas camp, Manila, Phillipine Islands, 36 months after internment (Brown, 1946). The corresponding means for the Germans can only be estimated approximately as the data for them are in decennial age groups and no finer age distribution for the American series is given. Also the German weights are standardized to heights which are probably rather below the unknown averages for the American series:

Mean weights (nude) in lb.

Ages		American	German	Difference
Men	19–40	124	128	4
	41–60	122	130	8
	Over 60	119	128	9
Women	19–40	101	113	12
	41–60	100	112	12
	Over 60	96	109	13

The differences here are substantial and greater for women than for men. The evidence considered in this paper is consistent in showing for all adult ages that loss of weight due to restricted rations is greater for females than for males.

This paper is chiefly concerned with a treatment of unpublished official records. The writer acknowledges with thanks permission to use such material given by: (*a*) the Public Health Branch of the Control Commission for Germany (British Element), by courtesy of Brigadier W. Strelley Martin, M.C., Public Health Adviser, Health Branch, C.C.G., and F. D. G. Bailey, Esq., formerly Nutrition Officer, Health Branch C.C.G.; (*b*) the Medical Directorate of the Air Ministry, by courtesy of Air Vice-Marshal P. C. Livingston, C.B., C.B.E., A.F.C., F.R.C.S., Director-General of Medical Services, Royal Air Force; (*c*) the Ministry of Food, by courtesy of the Chief Scientific Officer to the Ministry.

REFERENCES

BROWN, B. H. (1946). A detailed report on the weights and weight losses of twenty-four men in Santo Tomas internment camp. *J. Lab. Clin. Med.* **31**, 1129–32.

CATHCART, E. P., HUGHES, D. E. R. & CHALMERS, J. G. (1935). The physique of man in industry. *M.R.C., Industr. Heal. Res. Board, Report* no. 71.

CATHCART, E. P., BEDALE, E. M., BLAIR, C., MACLEOD, K. & WEATHERHEAD, M. (1927). The physique of women in industry. *M.R.C., Industr. Fat. Res. Board, Report* no. 44.

CONTROL COMMISSION FOR GERMANY (BRITISH ELEMENT), PUBLIC HEALTH BRANCH, NO. 1 NUTRITION SURVEY TEAM (1947). Report on a body-weight survey in the British zone.

COON, C. S. (1939). *The Races of Europe.* New York: Macmillan.

DAVIDSOHN, H. (1919). Die Wirkung der Aushungerung Deutschlands auf die Berliner Kinder mit besonderer Berücksichtigung der Waisenkinder der Stadt Berlin. *Z. Kinderheilk.* **21**, 349–407.

DUNSTAN, W. R. (1925). Height and weight of school children in an English rural area. *Metron*, **5**.

ELDERTON, E. M. (1914). Height and weight of school children in Glasgow. *Biometrika*, **10**, 288–339.

ELLIS, R. W. B. (1945). Growth and health of Belgian children during and after the German occupation (1940–1944). *Arch. Dis. Childhood*, **20**, 97–109.

FRIEND, G. E. (1935). *The Schoolboy: a Study of his Nutrition, Physical Development and Health.* Cambridge: Heffer.

FRIEND, G. E. & BRANSBY, E. R. (1947). Physique and growth of schoolboys. *Lancet*, 8 Nov., p. 677.

KEMSLEY, W. F. F. (1945). *A Preliminary Account of the Ministry of Food's Survey into the Weight and Height of the Population during* 1943. Ministry of Food Surveys Branch, DPD/W/6.

KEMSLEY, W. F. F. (1946). *Body Weight Survey: Report on the Weighings taken between April* 1943 *and October* 1945. Ministry of Food Surveys Branch, DPD/W/7.

KEMSLEY, W. F. F. (1947). *Body Weight Survey: Variations in Body Weight*—1943 *to* 1946. Ministry of Food Surveys Branch, DPD/W/8.

KOTELMANN, K. (1879). Die Körperverhältnisse der Gelehrtenschüler des Johanneums in Hamburg. *Z. kgl. preuss. stat. Bur., Berlin*, **19**, 1–16.

MINISTRY OF NATIONAL SERVICE, 1917–19 (1920). *Report, Vol.* 1, *upon the Physical Examination of Men of Military Age by National Service Medical Boards from November* 1*st*, 1917 *to October* 31*st*, 1918. H.M.S.O.

MORANT, G. M. (1943). Preliminary report on body measurements of 2,400 candidates for air-crew. *Flying Personnel Res. Comm.* no. 538.

MORANT, G. M. (1947). The heights and weights of Royal Naval pilots compared with those of R.A.F. pilots. *Flying Personnel Res. Comm.* no. 673.

MORANT, G. M. & GILSON, J. C. (1945). A report on a survey of body and clothing measurements of Royal Air Force personnel. *Flying Personnel Res. Comm.* no. 633*a*.

MUMFORD, A. A. (1927). *Healthy Growth: a Study of the Relation between the Mental and Physical Development of Adolescent Boys in a Public Day School.* Oxford University Press.

RANKE, O. (1905). Beiträge zur Frage kindlichen Wachstums: anthropologische Untersuchungen ausgeführt an holsteinischen Kindern von der Geburt bis zum vollendeter 15 Jahre. *Arch. Anthrop.* N.F., **3**, 161–80.

REID, D. D. (1947). Some measures of the effect of operational stress on bomber crews (F.P.R.C. 605). *Psychological Disorders in Flying Personnel of the Royal Air Force investigated during the War* 1939–1945, Air Publ. 3139, pp. 245–58. H.M.S.O.

RIETZ, E. (1903). Das Wachstum der Berliner Schulkinder während der Schuljahre. *Arch. Anthrop.*, N.F., **1**, 30–40.

SCHLESINGER, E. (1917). Das Wachstum der Knaben und Jünglinge vom 6 bis 20 Lebensjahr. *Z. Kinderheilk.* **16**, 265–304.

SCHMID-MONNARD, K. (1895). Über den Einfluss der Jahreszeit und der Schule auf das Wachstum der Kinder. *Jahr. Kinderheilk.* **40**, 84–107.

TESTING THE SIGNIFICANCE OF CORRELATION BETWEEN TIME SERIES*

By G. H. ORCUTT and S. F. JAMES, *Department of Applied Economics, University of Cambridge*

I. Introduction

Verification that two things are related to each other is achieved by showing empirically that there is a correspondence between their behaviours greater than could be expected by chance. Where continued experiment is not possible the data available are usually severely limited in both quantity and range of variation and it therefore becomes essential to have precise ideas whether the agreement with a hypothesis is actually a verification of it or whether it is more reasonable to consider the agreement as merely the result of a chance correspondence. This paper discusses the problem of deciding when a correlation between two economic time series is great enough to make it unreasonable to assume that the series are unrelated.

The testing of the significance of a correlation involves a comparison with what would have been obtained between non-related series thought to be analogous to the observed series. And, of course, the significance found for the correlation will depend upon the analogy deemed to be appropriate. The choice of an analogy depends upon experience as to which aspects of the real series being correlated are vital, in the sense that they affect the probability of obtaining chance correlations between non-related series. We can never be certain that some important aspect has not been overlooked, but, as our experience is broadened and we learn to take into account more and more factors, the chances of our running into a situation in which the analogy we choose is actually misleading becomes less and less. If we were forced to base our choice of an appropriate analogy to use in any individual situation on the data of that situation alone, the uncertainty of our tests of significance would be very great. Usually, however, the situation is one which we have learned to be similar to some larger class of experiences, and it is this larger class of experiences that furnishes a greater measure of information as to the analogy appropriate to all members of the class.

The most commonly used sampling model for generating series of independent terms is to draw them at random from a normal population of values; and in applying tests based upon this sampling model one must take account of the length of the series being dealt with. Fortunately, there is some evidence that tests of significance based on this sampling model are insensitive to variation of the frequency distribution of the population of values from which the random sampling is done (Pearson, 1931), and this makes it reasonable to apply such tests even when little is known of the frequency distributions from which the items of our real series have been drawn. There is, however, one obvious point at which the analogy underlying such tests of significance may break down when one is concerned with economic time series, namely, if the consecutive terms are really correlated. In economic time series, in meteorological time or spatial series, or, for that matter, in biological time series, autocorrelation usually exists. Production, or employment, or price-level series never go directly from

* Mr James was largely responsible for the part embodied in section III of this study. We both wish to express our appreciation for the large measure of assistance given us by Mr Richard Stone.

high values to low values, but, instead, high values are followed by values which are also high and a transition from high values to low values only takes place over a period of time. How closely successive values are related will of course partly depend upon the time between measurements and, as this time is made shorter and shorter, successive values of the series of measurements become more and more like their immediate neighbours. Autocorrelations are often very high and remain high even as the series are lengthened, whereas in random series the autocorrelations are small tending to zero as the series increase in length. In economics, most of the material that we wish to investigate for relationships exhibits autocorrelation, and there is a real need for a test of significance for correlations which is based on a more realistic sampling model.

Bartlett (1935) obtained the following large sample approximation of the variance of the sample correlations between two autocorrelated series having a true correlation of zero:

$$\text{var } r \sim \frac{n+2[(n-1)\rho_1\rho_1'+(n-2)\rho_1^2\rho_1'^2+\ldots+\rho_1^{n-1}\rho_1'^{n-1}]}{n^2}, \tag{1}$$

or more approximately

$$\text{var } r \sim \frac{1}{n}\frac{1+\rho_1\rho_1'}{1-\rho_1\rho_1'}, \tag{2}$$

where ρ_1 is the true value of the first autocorrelation of one of the two series and ρ_1' is that of the other. This is based on the assumption that each of the series was generated by the following type of process

$$x_t = \rho_1 x_{t-1} + \epsilon_t, \tag{3}$$

where the random error term, ϵ_t, is independent of x_{t-1} and $E(\epsilon_t) = 0$. Since then, Bartlett (1946, 1947) and Quenouille (1947) have given a large sample approximation of var r for correlations between any two autoregressive schemes having a true correlation of zero.* For linear autoregressive schemes it is

$$\text{var } r \sim \frac{n+2[(n-1)\rho_1\rho_1'+(n-2)\rho_2\rho_2'+\ldots+\rho_{n-1}\rho_{n-1}']}{n^2}, \tag{4}$$

or more approximately

$$\text{var } r \sim \sum_{t=-\infty}^{\infty} \rho_t\rho_t'/n. \tag{5}$$

Quenouille (1947) has given a convenient method of evaluating an expression such as (5).

Now while these formulae make it perfectly clear that, in interpreting the significance of a correlation between two series, it is necessary to take account of their autocorrelations, they have, as is recognized, certain practical limitations. Besides being based on the true autocorrelations, which are never obtainable in practice, they involve large sample assumptions which might not be reasonable for series of twenty or thirty items with which the economist must usually deal. It is also evident that, given only the formulae for var r, it is impossible, even neglecting the above considerations, to apply a test of significance without some knowledge concerning the shape of the distribution of sample correlations.

With the above difficulties in mind, we decided that a sampling experiment would give some guidance as to a reasonable procedure for carrying out tests of significance in the case of small samples. Since we could only carry out a rather limited sampling experiment, we were anxious that the sampling model used should generate unrelated series which were as analogous as possible to economic time series. In this way we might hope to obtain the

* The reader is also referred to a useful paper by Moran (1947), which gives the formula for the variance of the covariance between two series having known autoregressive properties.

maximum guidance in a region of practical interest. Yule (1921, 1926, 1927), Wold (1938) and Kendall (1944, 1945, 1946) have stressed that for most economic time series an autoregressive scheme is probably more relevant than the assumption of exact harmonic oscillation and each of the above has made considerable use of the linear second-order autoregressive scheme in studies of economic time series. Orcutt (1948*a*) tested the hypothesis that the economic time series used by Tinbergen (1939) might be considered to have been obtained by drawings from a single population of linear stochastic series all having the same underlying autoregressive structure. This hypothesis was brought to a test by comparing the means and variances at each lag of the correlograms of the economic series with the means and variances at corresponding lags of the correlograms of several sets of other series constructed according to a variety of models. On the basis of these comparisons and also a similar set of comparisons of correlograms of first differences, the conclusion was reached that so far as the evidence went the set of 52 economic series might have been obtained by drawings from the population of series generated by the model

$$Y_{(t+1)} = Y_t + 0{\cdot}3(Y_t - Y_{(t-1)}) + \epsilon_{(t+1)}, \tag{6}$$

where ϵ is random in time and has an expected value of zero.

Since equation (6) appears to us as the best available model for generating non-related series which are analogous to economic time series, we have used it as a basis for generating the series used in the remainder of this paper. It should be noted that equation (6) does not generate stationary series but rather a Brownian type of movement having no true mean. See Wold (1938, p. 53) for the distinction between stationary and evolutive time series. On the other hand, the series generated by equation (6) are not explosive in the sense that they tend to deviate from any given point or set-up oscillations of ever increasing amplitude. Six 30-item segments which were selected without regard to shape from a long series generated by equation (6) are shown in Fig. 1. Since the formulae given earlier for var r were derived on the assumption of stationary autoregressive processes, it is clear that on this account alone it would not be safe without additional evidence to apply them to correlations between non-stationary series such as generated by equation (6).

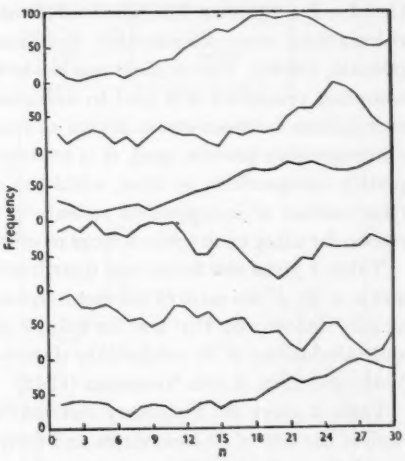

Fig. 1. 30-item segments of constructed series.

II. EMPIRICAL DISTRIBUTIONS OF CORRELATIONS BETWEEN NON-RELATED SERIES GENERATED BY THE AUTOGRESSIVE PROCESS

$$Y_{(t+1)} = Y_t + 0{\cdot}3(Y_t - Y_{(t-1)}) + \epsilon_{(t+1)}$$

We therefore sought to obtain empirically some idea of the distribution of correlations to be expected by chance between series drawn at random from the population of non-related series generated by the stochastic process of equation (6). In order to do this we first constructed a series of 3240 items by means of (6). The random elements used were two digit

numbers taken from Tables of Random Sampling Numbers by M. G. Kendall and B. Babington Smith (1939). They were read from left to right and double zeros were omitted. To obtain a true mean of zero, 50 was subtracted from each. Thus the random numbers used were drawn from a population having a rectangular distribution and a range of -49 through 0 to $+49$.

Our long series of 3240 items was first divided into 36 series each of 90 items. Each of these 36 series was then divided into three series of 30 items. The first 30-item series of the first 90-item series was labelled 1 A, the second 1 B and the third 1 C. The first 30-item series of the second 90-item series was labelled 2 A, the second 2 B and so on.

By the use of the usual product-moment formula for the correlation coefficient, correlations were obtained between all possible pairs of A series, between all possible pairs of B series and between all possible pairs of C series. That is, series 1 A was correlated in turn with series 2 A, 3 A, ..., 36 A; series 2 A was correlated in turn with series 3 A, 4 A, ..., 36 A, and so on. Thus we obtained 630 correlations between pairs of A series, 630 between pairs of B series and 630 between pairs of C series, making a total of 1890 correlations between different pairs of our 108 series each of 30 items. Then labelling the first 60 items of each 90-item series, series (A + B), we found the correlations between all possible pairs of the 36 (A + B) series. Having obtained these 630 correlations between series of 60 items we then found the 630 correlations obtained by correlating all the pairs of our 36 series of 90 items each. These series are labelled the (A + B + C) series. The labour of obtaining the above correlations was rather large but the calculations were considerably facilitated by use of a new type of calculating machine (Orcutt, 1948*b*). Now while it can be shown that, when the series are independent, the above sampling procedure will lead to unbiassed estimates of the moments of the population of correlations between series drawn at random from the universe of series generated by the autoregressive process used, it is evident that the 1890 correlations obtained are not completely independent, so that, while the effective number is substantially greater than 108 (the number of independent series), it, nevertheless, is considerably less than 1890. The reason for using each series a large number of times is simply the saving of labour.

Table 1 gives the frequency distribution for our constructed series with $n = 30$, $n = 60$, and $n = 90$, of the ratio of the mean square successive difference to the variance. This ratio is usually denoted by δ^2/s^2 and for infinite series $\delta^2/s^2 = 2(1-\rho_1)$. For a discussion of this ratio and tabulations of its probability distribution for random series, see von Neumann (1941, 1942) and Hart & von Neumann (1942).

Table 2 gives the frequency distributions of the correlations obtained between pairs of each of the sets of 30-item series and their total together with the frequency distributions of correlations obtained between pairs of the 60-item series and the 90-item series. Fig. 2 shows graphically the frequency distribution of the correlations for $n = 30$ together with a curve showing the frequency distribution of correlations between pairs of non-related random normal series with $n = 5$.* We tested the fit of this theoretical curve by means of a χ^2 test with 20 classes and 19 d.f., since the variance of the theoretical curve has been approximately fitted, and obtained a probability of less than 0·01. Figs. 3 and 4 show the distributions of the correlation coefficient which we obtained with $n = 60$ and $n = 90$, respectively. On the first we have drawn the curve for random series with $n = 5$ and on the second the

* The value 5 was chosen to make the variance $1/(n-1)$ agree as closely as possible with the observed variance 0·2736. Similarly $n = 6$ gives close agreement with the observed 0·2061 in Fig. 4.

curve for random series with $n = 6$. With 17 d.f., we obtained a χ^2 of 16·3 corresponding to a probability of about 0·5 in the first case and with the same degrees of freedom we obtained a χ^2 of 15·9 corresponding to a probability of about 0·7 in the second case. Table 3 gives the cumulative frequency distributions from the total of the 30-item series, and

Table 1. *Frequency distribution of δ^2/s^2 for series generated by* $Y_{(t+1)} = Y_t + 0{\cdot}3(Y_t - Y_{(t-1)}) + \varepsilon_{(t+1)}$

δ^2/s^2	Frequency		
	$n = 30$	$n = 60$	$n = 90$
0·00–0·10	40	23	30
0·10–0·20	27	9	6
0·20–0·30	23	1	0
0·30–0·40	5	1	0
0·40–0·50	4	1	0
0·50–0·60	4	1	0
0·60–0·70	2	0	0
0·70–0·80	2	0	0
0·80–0·90	1	0	0
Total	108	36	36
Mean δ^2/s^2	0·195	0·104	0·062

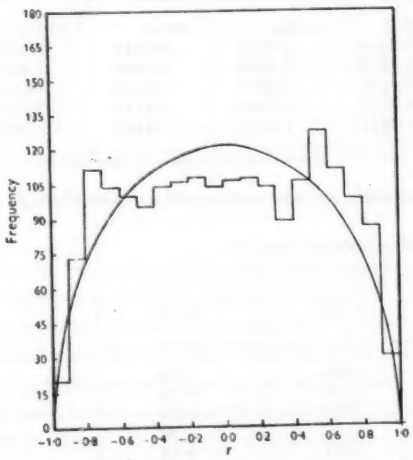

Fig. 2. Frequency distributions of r, $n = 30$.

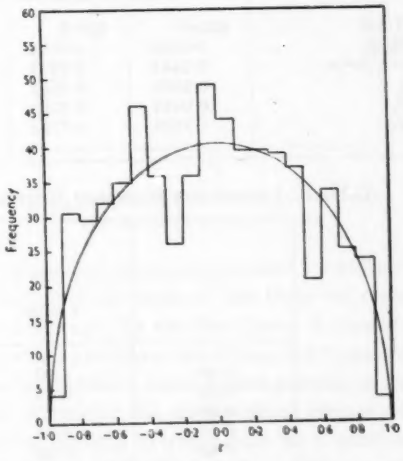

Fig. 3. Frequency distribution of r, $n = 60$.

for the 60- and 90-item series. The above application of the χ^2 test for testing the adequacy of the theoretical curves is admittedly rough, both because we have only approximately fitted the variances of the theoretical curves to our empirical distributions by our choice of n for the distributions of correlations between random series, and because the correlations making up our empirical distributions are not completely independent. The effect of both the above

Table 2. *Frequency distributions of r*

r	Frequency*					
	$n=30$				$n=60$ Series A+B	$n=90$ Series A+B+C
	Series A	Series B	Series C	Total A, B, C		
−1·0 to −0·9	6·5	12·0	2·0	20·5	4·0	3·5
−0·9 −0·8	28·0	32·5	12·0	72·5	30·5	10·5
−0·8 −0·7	35·0	49·0	27·5	111·5	29·5	25·0
−0·7 −0·6	37·0	37·5	29·0	103·5	32·0	31·0
−0·6 −0·5	33·0	31·0	36·0	100·0	35·0	37·0
−0·5 −0·4	37·5	25·5	32·0	95·0	46·0	40·0
−0·4 −0·3	43·5	18·5	41·5	103·5	36·0	38·5
−0·3 −0·2	36·0	32·0	38·0	106·0	26·5	37·0
−0·2 −0·1	30·0	31·5	47·0	108·5	36·0	44·5
−0·1 −0·0	34·0	47·0	22·5	103·5	49·0	45·5
+0·0 +0·1	31·0	39·0	36·5	106·5	44·0	54·0
+0·1 +0·2	30·0	31·5	46·0	107·5	39·5	46·0
+0·2 +0·3	34·5	26·5	42·5	103·5	39·5	30·5
+0·3 +0·4	29·5	25·0	34·0	88·5	39·0	46·5
+0·4 +0·5	30·0	28·0	48·0	106·0	37·0	38·5
+0·5 +0·6	40·5	44·5	43·5	128·5	21·0	31·5
+0·6 +0·7	36·5	39·0	35·5	111·0	33·5	27·5
+0·7 +0·8	37·0	30·0	31·0	98·0	25·0	30·0
+0·8 +0·9	25·0	37·0	24·0	86·0	23·5	11·5
+0·9 +1·0	15·5	13·0	1·5	30·0	3·5	1·5
Total	630·0	630·0	630·0	1890·0	630·0	630·0
Mean	0·0048	0·0004	0·0439	0·0164	−0·0248	−0·0023
$s^2=\Sigma r^2/n$	0·2844	0·3071	0·2293	0·2736	0·2446	0·2061
s	0·5333	0·5542	0·4788	0·5231	0·4945	0·4540
β_1	0·0047	0·0002	0·0335	0·0044	0·0140	0·0002
β_2	1·8125	1·7726	1·9238	1·8476	1·9423	2·0357

Table 3. *Cumulative frequency distributions with positive and negative r's combined*

$\lvert r\rvert$	Fraction greater than $\lvert r\rvert$		
	$n=30$	$n=60$	$n=90$
0·00	1·00	1·00	1·00
0·10	0·89	0·85	0·84
0·20	0·77	0·73	0·70
0·30	0·66	0·63	0·59
0·40	0·56	0·51	0·46
0·50	0·46	0·38	0·33
0·60	0·33	0·29	0·22
0·70	0·22	0·18	0·13
0·80	0·11	0·10	0·04
0·85	0·06	0·05	0·03
0·90	0·03	0·01	0·01
0·95	0·01	0·00	0·00
1·00	0·00	0·00	0·00

* Values of r were calculated to two places of decimals. Those on the border of two class-intervals were allocated one-half to each interval. The same applies to Tables 4 and 5.

circumstances will be to tend to exaggerate the χ^2's obtained and so underestimate the true probabilities that discrepancies between the empirical frequency distributions and the theoretical distributions are chance results.

Table 4 gives the distribution of 306 first-order partial correlations where $n = 30$ and two explanatory series are used. These 306 comprise the combinations available from our correlations between 30-item series using only partial correlations of the types, $r_{i(i+1).(i+2)}$, $r_{i(i+2).(i+1)}$ and $r_{(i+1)(i+2).i}$. Their frequency distribution is shown on Fig. 5 along with a curve showing the frequency distribution of correlations between pairs of non-related random series with $n = 5$.

Table 5 gives the distribution of 306 multiple correlations involving two explanatory series. These correlations were obtained from the above partial correlations and zero-order correlations. They are of the types, $R_{i.(i+1)(i+2)}$, $R_{(i+1).i(i+2)}$ and $R_{(i+2).i(i+1)}$. Their frequency distribution is shown on Fig. 6.

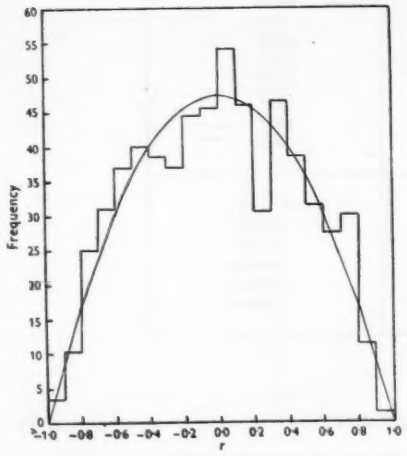

Fig. 4. Frequency distribution of r, $n = 90$.

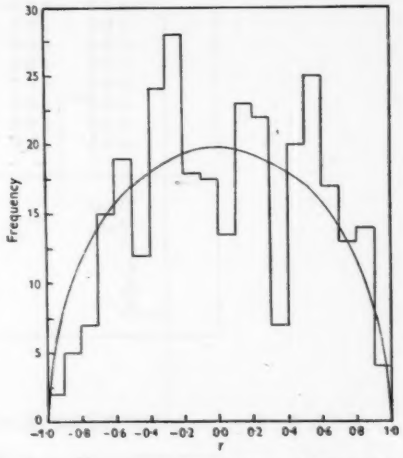

Fig. 5. Frequency distribution of first order partial correlations, $n = 30$.

In the next section we shall investigate whether sample-values can be used for unknown parameters in determining probabilities for the testing of correlations, but there are certain observations which might usefully be made at this stage. In the first place, it should be evident that the empirical distributions of this section provide a test of the null hypothesis that a given sample correlation might have occurred between series drawn at random from the population of series generated by equation (6). Since the distributions change very slowly with n, the length of the series, it should be possible to interpolate for a particular value of n with as much accuracy as our distributions justify. The position with regard to partial and multiple correlations is not so satisfactory since we have merely obtained, for $n = 30$, distributions of first-order partial correlations and multiple correlations involving two explanatory series. It is hoped, however, that these will be sufficient to give some idea of how high these coefficients must be in order to provide significant evidence against the null hypothesis and to throw some light on the possibility of getting very high multiple correlations between non-related series when n is less than 30 and four or five explanatory series are used.

Table 4. *Frequency distribution of first order partial correlations between series with* $n = 30$

$r_{ij.k}$	Frequency
−1·0 to −0·9	2·0
−0·9 −0·8	5·0
−0·8 −0·7	7·0
−0·7 −0·6	15·0
−0·6 −0·5	19·0
−0·5 −0·4	12·0
−0·4 −0·3	24·0
−0·3 −0·2	28·0
−0·2 −0·1	18·0
−0·1 −0·0	17·5
+0·0 +0·1	13·5
+0·1 +0·2	23·0
+0·2 +0·3	22·0
+0·3 +0·4	7·0
+0·4 +0·5	20·0
+0·5 +0·6	25·0
+0·6 +0·7	17·0
+0·7 +0·8	13·0
+0·8 +0·9	14·0
+0·9 +1·0	4·0
Total	306·0
Mean	0·0502
s^2	0·2341
s	0·4838
β_1	0·1053
β_2	1·9603

Table 5. *Frequency distribution of multiple correlations between series with* $n = 30$

$R_{i.jk}$	Frequency
+0·9 to +1·0	22·0
+0·8 +0·9	53·0
+0·7 +0·8	70·0
+0·6 +0·7	42·5
+0·5 +0·6	39·0
+0·4 +0·5	29·0
+0·3 +0·4	18·5
+0·2 +0·3	18·5
+0·1 +0·2	8·5
+0·0 +0·1	5·0
Total	306·0
Mean	0·6314
s^2	0·4470
s	0·6686
β_1	0·4474
β_2	2·8261

In the second place, when we first discovered that the variance of the distribution of the zero order correlation does not become substantially smaller as n is increased* from 30 to 60 and 90, we thought this implied that little was to be gained by use of greater lengths of time series in forming estimates of inter-relationship. This, however, does not necessarily follow, even if it be granted that economic time series have approximately the autoregressive properties of our constructed series. In particular, it does not follow if we imagine that we are dealing with a relation between such time series in which the error term in the relation is a random variable of constant expected variance over time. In this case, the variance of the related series will continue to grow with time since there is no true mean, but the variance of the error term will not. Therefore, as the series become longer, the variance of the error term will become a smaller and smaller fraction of the variance of the series being explained. This implies that the correlation coefficient will become higher and higher approaching unity as the series approach an infinite length. Thus, whilst almost as high correlations are to be

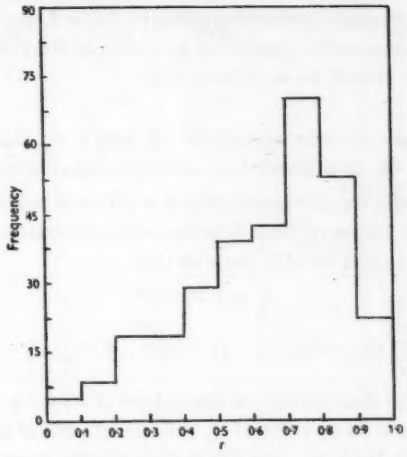

Fig. 6. Frequency distribution of $R_{t,5t}$, $n = 30$.

expected by chance between series of $n = 90$ as for $n = 30$, substantially higher correlations are to be expected as n increases if there is a real linear relation subject to a random error of constant expected variance over time. On the other hand, if the error term is not random in time but itself has continuity properties such as those of our series (6), then its expected variance will grow as the series become longer in the same way as the expected variance of our actual series grows. There will thus be no reason for the correlation to increase as the series become longer and we shall, in fact, be in the position of having gained little from the extension of the series. In this case consideration needs to be given to the possibility of correlating something like first differences in order to obtain any substantial advantage from the use of longer series.

In the third place, since for $n = 60$ and $n = 90$ we obtain good fits to our empirical distributions by means of the theoretical distributions of correlations between random series for

* We have received a very interesting letter from R. C. Geary in which he shows that for independent series generated by the rather similar process, $Y_t = Y_{t-1} + \epsilon_t$, the variance of correlations between series tends toward a non-zero finite positive quantity which cannot be very small.

$n = 5$ and $n = 6$, it follows that we might with some justification use these distributions for estimating the significance of correlations between economic time series. See David (1938) for tables of the correlation coefficient. Even in the case of $n = 30$, the use of the distribution for random series of $n = 5$ would not appear to overestimate the significance of correlations greater than 0·9.

Finally, it may be of some interest to note the result if we make use of the mean first lag autocorrelation of our 108 series with $n = 30$ and attempt to estimate var r by means of equation (1). However, instead of evaluating equation (1) as given, we reduced it to the following expression

$$\text{var } r \sim \frac{(1+r_1r_1')}{n(1-r_1r_1')} - \frac{2r_1r_1'(1-r_1^nr_1'^n)}{n^2(1-r_1r_1')^2}. \tag{7}$$

We have also substituted r_1 for ρ_1 and r_1' for ρ_1' since we intend to use sample rather than theoretical values of the autocorrelations. Then using $1-\frac{1}{2}$ (the mean δ^2/s^2 as given on Table 1) for r_1 and r_2 we obtain var $r = 0{\cdot}2759$. This estimate should be compared with the variance 0·2736 of our empirical distribution given in Table 2, and considering that Bartlett assumed very long stationary series generated by a simple Markoff scheme, it is remarkable that the estimate of var r should be as close as it is.

III. Investigation of the relation of var r to the sample values of the first lag autocorrelations

For the purpose of this study the 108 series, with $n = 30$, were separated into seven groups on the basis of the observed values of their first lag autocorrelation, r_1. The definition of the autocorrelation coefficient used for this purpose was

$$r_1 = 1 - \tfrac{1}{2}\delta^2/s^2, \tag{8}$$

where $$\delta^2 = \sum_1^{n-1} (Y_{i+1}-Y_i)^2/(n-1), \quad \text{and} \quad s^2 = \sum_1^{n} (Y_i-\bar{Y})^2/n.$$

As already mentioned, the distribution of the values of δ^2/s^2 for our 108 series is given in Table 1. The limits of r_1, the mean value of r_1, and the number of series for each of the seven groups of series is given in Table 6. The limits were chosen so as to obtain approximately equal numbers of series in each group.

Table 6. *Classification of series on the basis of sample values of first lag autocorrelations*

Group no.	Limits of the r_1 included in group	r_1 for group	No. of series in group
1	0·570–0·832	0·732	15
2	0·832–0·887	0·866	15
3	0·887–0·912	0·896	15
4	0·912–0·946	0·927	18
5	0·946–0·963	0·955	18
6	0·963–0·975	0·970	14
7	0·975–0·988	0·980	13

The set of 1890 correlations, for $n = 30$, were sorted into 28 classes corresponding to the combinations of groups from which the paired series yielding the correlations were drawn. The variance of these correlations about zero is given for each of the 28 classes by the top

number in each box of Table 7. The middle figure in each of these boxes is the number of correlations that occurred in the class. The bottom figure in each class is a 'theoretical variance' obtained by using equation (7) with r_1 and r'_1 taken as the mean values of the first lag sample autocorrelations corresponding to the groups of series being paired. For example, to estimate the variance of correlations between series from set 1 and from set 3, we used

Table 7. *Actual and theoretical values of* var *r for each of the* 28 *classifications of r based on the sample autocorrelations of the paired series**

Group no. of second series of each pair	Group no. of first series of each pair: 1	2	3	4	5	6	7
1	0·054 34 0·104	0·121 75 0·138	0·101 72 0·148	0·110 78 0·159	0·154 98 0·170	0·146 64 0·178	0·160 23 0·183
2		0·188 31 0·206	0·159 72 0·232	0·213 86 0·259	0·213 92 0·300	0·263 73 0·312	0·267 69 0·324
3			0·135 37 0·259	0·209 102 0·295	0·274 81 0·338	0·243 69 0·366	0·340 56 0·375
4				0·308 57 0·343	0·282 89 0·403	0·316 98 0·440	0·352 61 0·457
5					0·403 52 0·464	0·459 76 0·532	0·601 90 0·632
6						0·510 31 0·616	0·599 49 0·637
7							0·784 27 0·701

* The top number in each box is the actual variance, the middle number is the number of correlations in the class, and the lower number in each box is a 'theoretical' variance.

0·732 for r_1 in equation (7) and 0·896 for r'_1. It will be noticed that equation (7) only differs from the more approximate form of equation (2) in the right-hand term. For low values of r_1 or r'_1 or large n, this term will be insignificant. Thus, whereas var r, as estimated by equation (7) for class 1–1, is only 0·006 less than the same estimate made by equation (2), we find that for class 4–4 the use of equation (7) gives an estimate which is 0·097 less than that obtained using equation (2). For higher values of $r_1 r'_1$ the difference becomes rapidly greater.

In Fig. 7 we have plotted the observed value of the variance for each class against the 'theoretical' value as obtained using equation (7). The straight line on this diagram is merely the 45° slope and would hold if the 'theoretical' variance agreed exactly with the actual. Notwithstanding some evidence of a systematic disagreement, the agreement obtained is remarkably good when it is considered that the assumptions under which Bartlett obtained the formula used for var r are far from being realized in this case. Not only are we dealing with non-stationary series having a small n, but we have used sample autocorrelations rather than true autocorrelations, and in addition our series were not generated by a Markoff process. Evidently we shall not go far wrong, even with our type of series, if for purposes of testing the null hypothesis we estimate the variance of a correlation between two series by using equation (7). The error that this involves for series of our type would appear, on the basis of Fig. 7, to be an overestimation of the variance to the extent of about 15 % for almost the entire range of sample first lag autocorrelations covered by our experiment.

Table 8. *Frequency distributions of the absolute values of correlations between series grouped according to their sample autocorrelations*

$\lvert r \rvert$	A	B	C	D	E	F	G
	Classes 1–1, 1–2, 1–3, 1–4	Classes 1–5, 1–6, 1–7, 2–2	Classes 2–3, 2–4, 3–3, 3–4	Classes 2–5, 2–6, 2–7, 3–5	Classes 4–4, 3–6, 3–7, 4–5	Classes 4–6, 4–7, 5–5, 5–6	Classes 6–6, 5–7, 6–7, 7–7
0 –0·10	56·50	56·25	37·00	29·75	14·50	19·75	1·00
0·10–0·20	60·25	37·00	45·00	33·50	20·50	21·75	2·25
0·20–0·30	45·75	43·50	40·50	29·25	32·00	19·50	4·00
0·30–0·40	30·75	35·50	41·50	34·00	30·75	17·75	3·00
0·40–0·50	28·00	27·25	41·50	49·50	28·75	24·00	5·25
0·50–0·60	21·50	25·50	37·25	48·50	47·50	32·00	12·00
0·60–0·70	12·25	18·00	30·25	38·25	44·75	37·75	24·75
0·70–0·80	4·00	19·25	15·00	36·50	32·75	61·00	44·25
0·80–0·90	0·00	3·75	9·00	15·75	17·75	48·00	58·25
0·90–0·95	0·00	0·00	0·00	0·00	1·75	5·50	31·25
0·95–1·00	0·00	0·00	0·00	0·00	0·00	0·00	11·00
Sum	259	266	297	315	271	287	197
Mean-square about 0	0·1070	0·1595	0·1889	0·2527	0·2885	0·3776	0·6138
β_2 (using moments about zero)	2·447	2·283	1·619	1·674	1·559	1·462	1·140

Before the above method for estimating var r can be fully utilized, it is necessary to know something about the shape of the distributions involved and, in order to provide information about this point, we obtained the frequency distributions tabulated in Table 8 and shown in Figs. 8–14. The 28 classes of correlations were ordered according to the theoretical values of var r as given in Table 7. Then the correlations of the first four classes were grouped together into class A, the second four classes were grouped together into class B, and so on. Because these distributions should be symmetrical about zero and in view of the small sample sizes, we have obtained frequency distributions of absolute values of r. These distributions make it clear that the distribution of r for fixed values of ρ_1 and ρ_1' depends on r_1 and r_1' to a very

considerable extent. It is rather interesting to note the way in which the mode gradually moves from zero towards 1 as the variance of the distribution increases. It is also interesting to note that, even for very high variances such as for distributions F and G, the ordinates of the curves still appear to approach zero for r approaching unity.

We were interested in seeing how well our empirical frequency distributions could be fitted by frequency distributions of approximately the same variance of correlations between series of normally distributed random items. Since the variance of correlations between random series is $1/(n-1)$, we have used the distribution for random series of $n = 11$ to fit the frequency distribution of A, Fig. 10, the distribution for random series of $n = 7$ for B, that of $n = 6$ for C, and that of $n = 5$ for D. We did not bother with the rest since it is obvious that the fit would be completely unsatisfactory. As a rough test of the goodness of the above fits, we applied a χ^2 test in each case. In case A we obtained a χ^2 of 2·663 with 6 d.f. and this corresponds to a probability of above 0·8. In case B, χ^2 was 5·105 with 6 d.f. and this corresponds to a probability of about 0·5. For C, χ^2 was 9·082 with 8 d.f. and this corresponds to a probability of about 0·3. In case D, χ^2 was 33·634 with 8 d.f. and this has a probability of less than 0·001. The same remarks apply concerning the roughness of the above test as we made in § II and it should also be true here that our use of the χ^2 test tends to underestimate the probabilities.

The primary implication of the results given in this section appears to us as follows. A reasonable way of testing the significance on the null hypothesis of a correlation between economic time series is first to estimate var r by means of equation (7) and, in so doing, to use the sample values of the first lag autocorrelations of the two series. Secondly, if the estimated var r is less than about 0·25, then insert it into the formula var $r = 1/(n'-1)$ and evaluate n'. Round n' off to the nearest integer and make use of the distribution of r that applies for random series with n equal to this integer to evaluate the probability that r might have occurred by chance between two non-related series. If the estimated var r turns out to be more than 0·25, then it seems unlikely that this method is valid, but the above test of significance would appear to be stronger than necessary and so would at least be a safe one to apply.

Secondly, our study would seem to offer some evidence that the variance of correlations between pairs of unrelated series of a given n, and with given sample autocorrelations, will be nearly if not completely independent of the true autocorrelations of the series. This follows from the fact that in Fig. 9 the points follow very closely the line that they would have been expected to follow if we had been dealing with sets of correlations associated with sets of series having different true autocorrelations. If the true autocorrelations were actually known, one might, at least in principle, first evaluate the probability of independently drawing two series with certain sample autocorrelations, and secondly evaluate the probability of obtaining by chance a given correlation between series having their sample autocorrelations. Having done this, it might be possible to evaluate the combined probability of drawing two series independently which had the autocorrelations and correlation between them which were actually obtained.

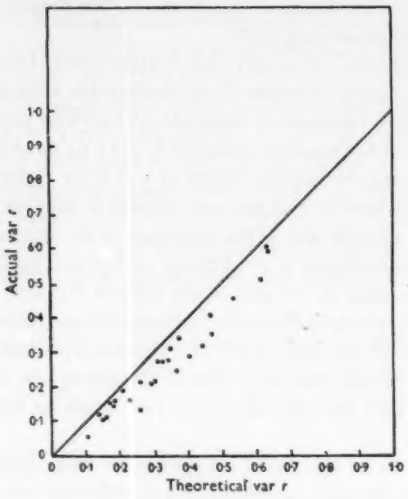

Fig. 7.

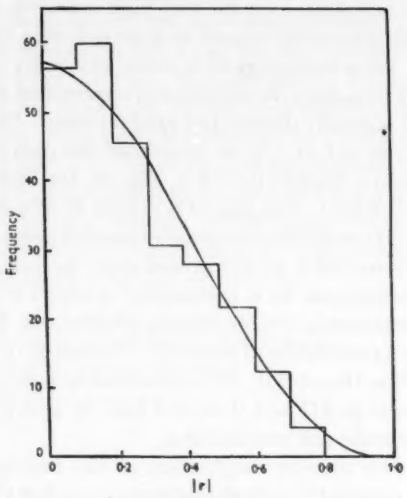

Fig. 8A.

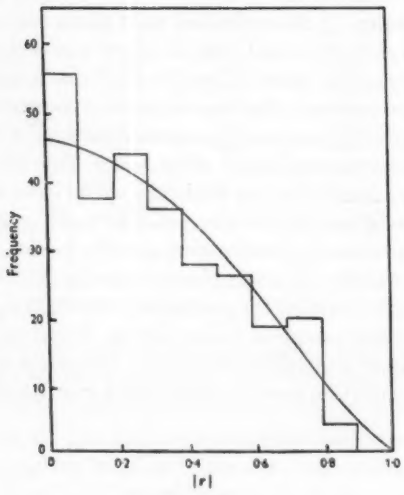

Fig. 9B.

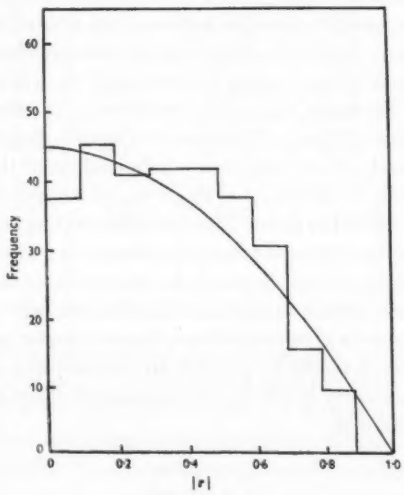

Fig. 10C.

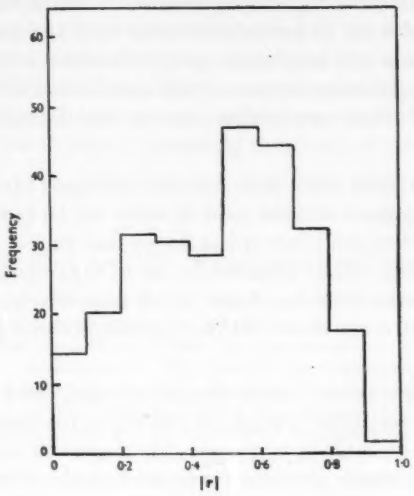

Fig. 11D.

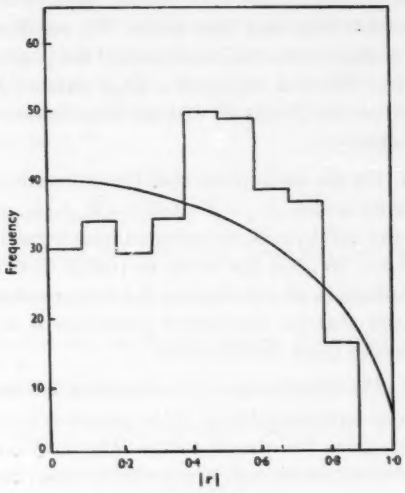

Fig. 12E.

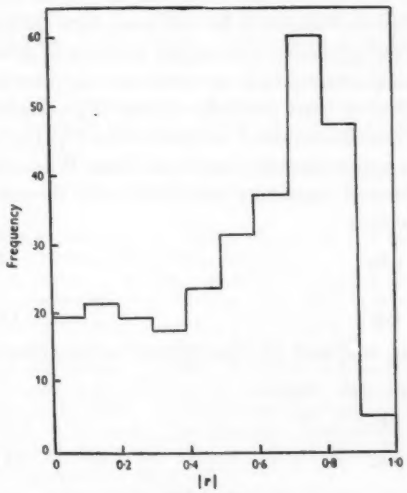

Fig. 13F.

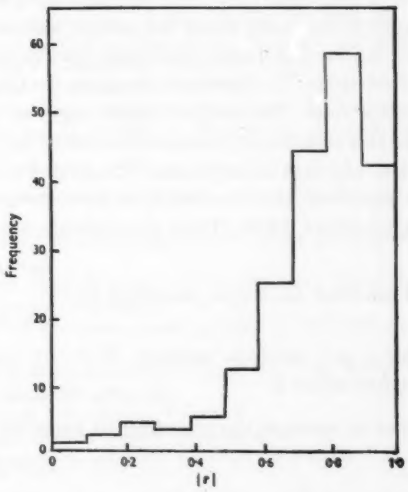

Fig. 14G.

IV. Summary and a general remark on the problem of detecting real relations

The problem dealt with in this paper was that of testing the significance of correlations between economic time series. We constructed a set of non-related series with the same properties as are believed to hold for yearly series of a substantial group of economic time series. We then obtained a large number of correlations between these constructed series and on the basis of various distributions of these correlations came to the following conclusions.

1. On the assumption that the economic time series being dealt with are analogous to our specific model, $Y_{t+1} = Y_t + 0{\cdot}3(Y_t - Y_{t-1}) + \epsilon_t$, correlations between pairs of series can be tested for the null hypothesis using our empirical distributions. Since it was shown that for $n = 60$ and $n = 90$ good fits to our empirical distributions can be obtained by use of the ordinary distribution of correlations between random series with $n = 5$ and $n = 6$ respectively, it follows that an alternative procedure is to test correlations between economic series by means of these distributions.

2. The distribution of correlations between non-related series depends primarily on the sample autocorrelations of the paired series and very little, if at all, on the true autocorrelations, given the sample values. This makes it reasonable to apply tests based upon sample autocorrelations, and it was shown that one reasonable procedure is to use equation (7) to estimate the variance of r and then use the estimated variance of r to select the appropriate distribution of correlations between random series. Having chosen this distribution, one can then test the significance of the correlation in the usual way. The properties of this conditional test might repay a theoretical examination.

3. If economic time series are analogous to the constructed series used in this paper then, except in the cases where the sample autocorrelations happen to be low, such high correlations between economic time series may be expected by chance that we are unlikely to detect real relations. The distributions given for the partial and multiple correlations only accentuate this view. One method which suggests itself of at least partially extricating ourselves from this situation is to make an autoregressive transformation of the time series involved in such a way that at least one of the series becomes approximately random in time. When this has been done, the correlation between the transformed series may then be tested in the usual way (Bartlett, 1935). Thus, for example, suppose that

$$x_{1t} = \beta x_{2t} + u_t, \tag{9}$$

and the error term u_t is generated by

$$u_t = \alpha u_{t-1} + \epsilon_t, \tag{10}$$

where ϵ_t is a random variable. If $\beta = 0$, then $x_{1t} = u_t$ and an appropriate autoregressive transformation is $x'_{1t} = x_{1t} - \alpha x_{1(t-1)}$, $x'_{2t} = x_{2t} - \alpha x_{2(t-1)}$,

so that in terms of the transformed variables we have

$$x'_{1t} = \beta x'_{2t} + \epsilon_t. \tag{11}$$

Since one of the two variables is now random we can apply the usual test of significance of correlation between x'_{1t} and x'_{2t} with the consequent advantage of a great increase in the effective degrees of freedom. On the other hand, if $\beta \neq 0$, then the expected value of the least square estimate of β in equation (11) will still be the same as in equation (9) and the expected

value of estimates of the true correlation between x'_{1t} and x'_{2t} will still be very near'y the same as the expected value of estimates of correlation between x_{1t} and x_{2t}. This means that in this case, at least, the transformed form (11) is far more effective than the untransformed form (9), for the purpose of testing the null hypothesis. When $\beta \neq 0$, then the appropriate autoregressive transformation for estimating β is not one which leaves one of the series random but rather one that leaves the error term random. However, in the case $\beta = 0$, it is evident that when the error term is random then the variable which is taken to be dependent must also be random since it is entirely composed of the error term.*

REFERENCES

AITKEN, A. C. (1935). On least squares and linear combination of observations. *Proc. Roy. Soc. Edinb.* **55**, 42.

BARTLETT, M. S. (1935). Some aspects of the time correlation problem in regard to tests of significance. *J. Roy. Statist. Soc.* **98**, 536.

BARTLETT, M. S. (1946). On the theoretical specification and sampling properties of autocorrelated time-series. *J. Roy. Statist. Soc.* Supplement, **8**, 27.

BARTLETT, M. S. (1947). *Stochastic Processes.* Mimeographed notes of a course given at the University of North Carolina in the Fall Quarter, 1946.

CHAMPERNOWNE, D. C. (1948). Sampling theory applied to autoregressive time-series. *J. Roy. Statist. Soc., Series B*, **10** (in press).

COCHRANE, D. & ORCUTT, G. H. (1949). Application of least squares regression analysis to relationships containing autocorrelated error terms. *J. Amer. Statist. Ass.* **44** (in press).

DAVID, F. N. (1938). *Tables of the Correlation Coefficient.* Cambridge University Press.

HART, B. I. & VON NEUMANN, J. (1942). Tabulation of the probabilities for the ratio of the mean square successive difference to the variance. *Ann. Math. Statist.* **13**, 207.

KENDALL, M. G. & BABINGTON SMITH, B. (1939). *Tracts for Computers*, no. XXIV. Cambridge University Press.

KENDALL, M. G. (1944). Oscillatory movements in English agriculture. *J. Roy. Statist. Soc.* **106**, 91.

KENDALL, M. G. (1945). On the analysis of oscillatory time-series. *J. Roy. Statist. Soc.* **108**, 93.

KENDALL, M. G. (1946). *Contributions to the Study of Oscillatory Time-Series.* National Institute of Economic and Social Research, Occasional Papers X. Cambridge University Press.

MORAN, P. A. P. (1947). Some theorems on time series. I. *Biometrika*, **34**, 281.

ORCUTT, G. H. (1948*a*). A study of the autoregressive nature of the time series used for Tinbergen's model of the economic system of the United States, 1919–1932. *J. Roy. Statist. Soc., Series B*, **10**, 1.

ORCUTT, G. H. (1948*b*). A new regression analyser. *J. Roy. Statist. Soc., Series A*, **111** (in press).

PEARSON, E. S. (1931). The test of significance for the correlation coefficient. *J. Amer. Statist. Ass.* **26**, 128.

QUENOUILLE, M. H. (1947). Notes on the calculation of autocorrelations of linear autoregressive schemes. *Biometrika*, **34**, 365.

TINBERGEN, J. (1939). *Statistical Testing of Business-Cycle Theories. Vol. 2: Business Cycles in the United States of America* 1919–1932. Geneva; League of Nations.

VON NEUMANN, J. (1941). Distribution of the ratio of the mean square successive difference to the variance. *Ann. Math. Statist.* **12**, 367.

VON NEUMANN, J. (1942). A further remark on the distribution of the ratio of the mean square successive difference to the variance. *Ann. Math. Statist.* **13**, 86.

WOLD, H. (1938). *A Study in the Analysis of Stationary Time Series.* Uppsala.

YULE, G. U. (1921). On the time correlation problem. *J. Roy. Statist. Soc.* **84**, 497.

YULE, G. U. (1926). Why do we sometimes get nonsense correlations between time series? *J. Roy. Statist. Soc.* **89**, 1.

YULE, G. U. (1927). On a method of investigating periodicities in disturbed series with special reference to Wolfer's sunspot numbers. *Philos. Trans.* A, **226**, 267.

* For a fuller discussion of the problem of estimation when the error term is autocorrelated, see Aitken (1935), Champernowne (1948) and Cochrane & Orcutt (1949).

MISCELLANEA

Note on the median of a multivariate distribution

By J. B. S. HALDANE

The median of a univariate distribution is an exceedingly useful parameter but, whereas the notions of the mean and mode can be applied without ambiguity to distributions in two or more dimensions, this is not so for the median. It is the object of this note to point out that when we are dealing with multivariate distributions, there are two quite distinct sets of parameters, each of which possess some of the properties of the univariate median, while lacking others.

The possibility arises from the fact that the univariate median is a location parameter associated with two quite different scale parameters. In the first place, for the distribution $dF = f(x)\,dx$, the median is defined as M, where

$$\int_{-\infty}^{M} dF = \int_{M}^{\infty} dF = \tfrac{1}{2}.$$

Integration here and throughout is understood in Stieltjes's sense.

When so defined, the median is obviously associated with the quartiles defined by $\int_{-\infty}^{Q_1} dF = \frac{1}{4}$ and $\int_{Q_3}^{\infty} dF = \frac{1}{4}$, and with the interquartile range $Q_3 - Q_1$.

Secondly, however, we may define the median as the value M which minimizes $\int_{-\infty}^{\infty} |M-x|\,dF$.

Similarly, the mean can be defined as the value m which minimizes $\int_{-\infty}^{\infty} (m-x)^2\,dF$. Now the minimum value of this quantity is simply the variance. Just as the mean is associated with the standard deviation as a measure of dispersion, so on this definition the median is associated with the mean deviation about the median. The more commonly used measure, the mean deviation about the mean, has perhaps less to recommend it, since it is not a stationary value, and therefore more liable to error if the corresponding scale parameter is in error. In geometrical language the median is the point the sum of whose distances from the representative points of the sample is a minimum.

Both these definitions of the median are equivalent in the univariate case, and both are of course indeterminate if the number of members of a sample is even, unless an even number of them coincide with the median. The various devices which avoid this indeterminacy represent the median as a limit.

When we pass to two or more dimensions these two definitions are no longer equivalent, and it seems worth while to distinguish the two analogues of the univariate median as the arithmetic and geometric medians.

If we have a number of variates $x, y, z, \ldots$ the arithmetic median is the set of values $(X, Y, Z, \ldots)$, where $X, Y, Z, \ldots$ are the medians of $x, y, z, \ldots$ defined in either of the two above ways. When x, y, z, etc., are different in kind it is obviously the only reasonable generalization. It has the merit of being invariant, like the median, when any of the variates is replaced by a monotonic function of it. But it is not invariant under a rotation of axes.

For consider the arithmetical median of three coplanar points. If we take rectangular axes their co-ordinates are (x_1, y_1), (x_2, y_2), (x_3, y_3). Those of the median are (x_m, y_m), where x_m is the middle value of x_1, x_2, x_3 if they are all different, and the value of the two equal ones if two are equal. Hence as we rotate the axes we find that the position of the arithmetic median changes. Unless it coincides with one of the apices of the triangle, one of the sides must subtend a right angle at it. In fact, the locus consists of those arcs of the circles which have the sides of the triangle as diameters which lie within the triangle (see Fig. 1). If the triangle has a right or obtuse angle, this angle lies on the locus, as does the foot of the perpendicular from it on the opposite side. If the triangle is acute angled, it passes through the feet of the three perpendiculars. Similarly, the locus of the arithmetic median of the vertices of a tetrahedron consists of portions of spheres. For an odd number of more than three points in a plane, the arithmetic median may always be one of them, or its locus may consist of a series of circular arcs. For an even number it is of course indeterminate, unless one or other of the special conventions devised for the univariate case is used.

The geometric median is defined as the point such that the sum of its distances from the sample points is a minimum. It is invariant under a change of axes, but is not invariant when the scales in different directions are altered. Its sole value is therefore in problems of geometrical probability. It occurred in a problem of this type during the recent war, and might perhaps be of value in studies on such aggregates as star clusters, where we desire to find a representative point which is less affected than the centroid by outliers which may not be members of the cluster.

The geometrical median of three coplanar points is the point in the triangle formed by them at which each side subtends an angle of $\frac{2}{3}\pi$, provided that no angle of the triangle exceeds $\frac{2}{3}\pi$. If one angle exceeds this value, the geometrical median is the obtuse vertex of the triangle. I have been unable to find any simple geometrical construction in the case of more than three points. It is, however, easy to show that

Fig. 1.

the geometrical median is unambiguously defined. For let us take it (or *per impossible*, one of the geometrical medians) as our origin of Cartesian co-ordinates. Consider a set of n coplanar points (x_r, y_r). First suppose that no sample point coincides with the origin, and if necessary rotate the axes so that no axis passes through a sample point. Let R be the sum of the distances of the sample points from the point $(x, 0)$. Then

$$R = \sum_{r=1}^{n} [(x-x_r)^2+y_r^2]^{\frac{1}{2}}, \quad \frac{dR}{dx} = \sum_{r=1}^{n} [(x-x_r)\{(x-x_r)^2+y_r^2\}^{-\frac{1}{2}}].$$

This must be zero when $x = 0$. But

$$\frac{d^2R}{dx^2} = \sum_{r=1}^{n} [y_r^2\{(x-x_r)^2+y_r^2\}^{-\frac{3}{2}}].$$

All the terms in this sum are necessarily positive, since the denominator is the cube of a distance which is taken as positive and can never change its sign. Hence d^2R/dx^2 is always positive, and R has only one minimum.

Next suppose that the median coincides with one of the points, say the first; then

$$R = |x| + \sum_{r=2}^{n} [(x-x_r)^2+y_r^2]^{\frac{1}{2}}, \quad \frac{dR}{dx} = \pm 1 + \sum_{r=2}^{n} [(x-x_r)\{(x-x_r)^2+y_r^2\}^{-\frac{1}{2}}],$$

d^2R/dx^2 is positive as before, but dR/dx has a saltus at $x = 0$, increasing in value by 2, and changing sign. R has therefore a sharp minimum. The proof in three or more dimensions is analogous. Changing to polar co-ordinates with the origin as centre and the co-ordinates of the sample points as (ρ_r, θ_r), it follows that if the median does not coincide with one of them,

$$\Sigma \cos\theta_r = \Sigma \sin\theta_r = 0,$$

whilst if it does so, these sums lie between ± 1. If several sample points coincide with the geometric median, the modifications are obvious.

It is clear that the minimum sum of the distances, divided by the number of points, is the many-dimensional analogue of the mean deviation from the median in one dimension.

To sum up, the arithmetical median is obviously to be preferred in ordinary statistical work, but the geometrical median has certain advantages in problems of geometrical probability. In either case it is desirable to state clearly how the median is defined.

A property of rank correlations

By H. E. DANIELS

The product-moment correlation coefficient may be considered a satisfactory measure of the association between two variates, both because of its special relevance to the bivariate normal distribution, and, more generally, from the fact that the square of its population value is the fraction of the variance of one variate removed by linear regression on the other.

When the data are presented in ranked form, however, the suitable choice of a measure of concordance between rankings is less obvious. Spearman's analogous use of the product-moment correlation coefficient ρ between the ranks cannot be so easily justified, though when calculated from the sum of squares of rank differences it is seen to measure in a rather arbitrary sense the degree of agreement between ranks. Kendall's coefficient τ has a more direct interpretation in that it is a function of the total number of corresponding pairs of ranks which agree in order, and Moran (1948) has further shown that τ is simply related to the least number of interchanges required to bring two rankings into perfect agreement. It may be said, therefore, that τ is a satisfactory measure of concordance in the sense that increasing values of τ correspond to increasing agreement between the rankings as defined in either of these ways.

In a previous paper (1944) I introduced the quantity

$$\Gamma = \frac{\Sigma a_{ij} b_{ij}}{\sqrt{(\Sigma a_{ij}^2 \, \Sigma b_{ij}^2)}}$$

as a general measure of the correlation between two sets of observations, ranked or otherwise, where a_{ij}, b_{ij} are scores assigned to corresponding pairs i, j of the two variables and $a_{ij} = -a_{ji}$, $b_{ij} = -b_{ji}$. Both ρ and τ are included as special cases. It is of interest to see how far Γ may be justified as a suitable measure of rank correlation.

Provided the scores have the same sign as the difference of the corresponding ranks, Γ takes the values ± 1 when there is respectively complete concordance or discordance between the rankings, and when the ranked attributes are independent, Γ is on the average zero (though it must be remembered that a zero average value of Γ does not necessarily imply independence). A further property is, however, required before Γ can be accepted as satisfactory; it has to be shown that increasing values of Γ correspond in some way to increasing concordance between the rankings.

A property of this type which one might expect of a rank correlation coefficient is that if any two corresponding pairs of ranks do not agree in order, the value of the coefficient should increase when the members of one of the pairs are interchanged. It is now shown that Γ has this property provided the scores a_{ij}, b_{ij} do not decrease with increasing rank separation and are not zero except for tied ranks. The scores for both ρ and τ satisfy the conditions.

Let p_i, q_i be the ranks of the ith members of the two sets of observations, and suppose that $p_r > p_s$, $q_r < q_s$, and that p_r, p_s are to be interchanged. The denominator of Γ is unaffected by any relative permutation of the ranks. Initially, the numerator is

$$\begin{aligned}\Sigma a_{ij} b_{ij} &= \Sigma'' a_{ij} b_{ij} + \Sigma'' a_{rj} b_{rj} + \Sigma'' a_{ir} b_{ir} + \Sigma'' a_{sj} b_{sj} + \Sigma'' a_{is} b_{is} + a_{rs} b_{rs} + a_{sr} b_{sr} \\ &= \Sigma'' a_{ij} b_{ij} + 2\Sigma'' a_{rj} b_{rj} + 2\Sigma'' a_{sj} b_{sj} + 2a_{rs} b_{rs},\end{aligned}$$

where Σ'' denotes summation over all values of i, j excluding r and s. After interchanging p_r, p_s it becomes

$$\Sigma'' a_{ij} b_{ij} + 2\Sigma'' a_{sj} b_{rj} + 2\Sigma'' a_{rj} b_{sj} - 2a_{rs} b_{rs},$$

which is an increase of

$$-2\Sigma''(a_{sj} - a_{rj})(b_{sj} - b_{rj}) - 4a_{rs} b_{rs} = -2\Sigma(a_{sj} - a_{rj})(b_{sj} - b_{rj}).$$

Now introduce the condition that a_{ij} is a non-decreasing function of $p_j - p_i$, and consider $a_{sj} - a_{rj}$ for all values of j. The following are the possible alternatives.

(i) $p_j \geqslant p_r > p_s$. $a_{sj} > 0$, $a_{rj} \geqslant 0$ and $a_{sj} \geqslant a_{rj}$.
(ii) $p_r > p_j \geqslant p_s$. $a_{sj} \geqslant 0$, $a_{rj} < 0$, so that $a_{sj} - a_{rj} = a_{sj} + a_{jr} > 0$.
(iii) $p_r > p_s > p_j$. $a_{sj} < 0$, $a_{rj} < 0$ and $a_{jr} \geqslant a_{js}$, so that $a_{sj} \geqslant a_{rj}$.

Thus in all cases $a_{sj} - a_{rj} \geqslant 0$ and in at least one case $a_{sj} - a_{rj} > 0$.

Similarly, since $q_r < q_s$ the fact that b_{ij} is a non-decreasing function of $q_j - q_i$ implies that $b_{sj} - b_{rj} \leqslant 0$ and < 0 at least when $a_{sj} - a_{rj} > 0$. It follows that Γ is increased by the interchange.

On the other hand, if the scores decrease with increasing rank separation the result is not necessarily true. For suppose $a_{ij} = \pm A$, a large number, when $p_j - p_i = \pm 1$, and $a_{ij} = \pm 1$ as for Kendall's τ when $|p_j - p_i| > 1$, with similar scores b_{ij} for $q_j - q_i$. Then for the ranking

8	6	7	1	4	2	5	3
1	2	3	4	5	6	7	8

the value of Γ is decreased on interchanging 6 and 4 in the first ranking.

When tied ranks are present, any pair of ties is scored zero, but some consideration will show that Γ is still increased if a discordant pair is brought to concordance by an interchange.

The interchange of a particular pair of discordant ranks will in general alter the order of some of the other pairs which involve one or other of the ranks interchanged, but it is worth remarking that by virtue of the result just proved the value of τ, and hence the total number of concordances between pairs of ranks, must be increased.

REFERENCES

MORAN, P. A. P. (1948). *Proc. Camb. Phil. Soc.* **44**, 142.
DANIELS, H. E. (1944). *Biometrika*, **33**, 129.

Approximation errors in distributions of independent variates

BY H. O. HARTLEY

1. Let x and y be two independent variates and let $q = \phi(x, y)$ be a function of these variates (e.g. the ratio x/y or the product $x.y$). In distribution theory one is frequently faced with the following problem:

If we approximate to the distribution of x and/or y, what is the effect of such approximations on the distribution of q?

In this note we derive a lemma which, although mathematically trivial, provides a gauge for this error and has been of great help in recent work. It appeared worth while therefore to put it on record (see § 5).

2. To fix the ideas we assume that $0 < x < \infty$, $0 < y < \infty$, and that the differentiable function $q = \phi(x, y)$ is monotonic increasing in x and monotonic in y, so that differentiable and monotonic inversion functions $x = \psi(q, y)$ and $y = \chi(q, x)$ exist. It is obvious from the argument given below that some of these restrictions are in fact not essential.

3. Let $F(X)$ be the chance for $0 \leqslant x \leqslant X$ and $G(Y)$ be the chance for $0 \leqslant y \leqslant Y$, and denote by $f(X)$ and $g(Y)$ approximations to F and G. We shall assume that these approximations are themselves 'probability integrals', so that

$$\lim_{X \to \infty} f(X) = \lim_{Y \to \infty} g(Y) = 1 \quad \text{and} \quad f'(X) \geqslant 0, \quad g'(Y) \geqslant 0, \tag{1}$$

where the distribution functions f' and g' are the differentials of f and g. Let us denote the differences between exact probability integrals and the approximations by

$$\epsilon(X) = F(X) - f(X), \quad \eta(Y) = G(Y) - g(Y), \tag{2}$$

so that

$$\lim_{X \to \infty} \epsilon(X) = \lim_{Y \to \infty} \eta(Y) = 0, \quad \epsilon(0) = \eta(0) = 0. \tag{3}$$

The probability integral of q, i.e. the chance ($H(Q)$ say) of $q = \phi(x, y) \leqslant Q$, is then given by

$$H(Q) = \int_0^\infty \int_0^\infty F'(x)\, G'(y)\, dx dy, \tag{4}$$
$$\phi(x, y) \leqslant Q$$

where F' and G' are the distribution functions of x and y. We have from (2) and (4)

$$\left.\begin{aligned} H(Q) &= \int_0^\infty \int_0^\infty f'(x)\, g'(y)\, dx\, dy + \int_0^\infty \int_0^\infty \{\eta' + g'\epsilon'\}\, dx\, dy \\ &\qquad \phi(x,y) \leqslant Q \qquad\qquad \phi(x,y) \leqslant Q \\ &= h(Q) \qquad\qquad + \theta(Q) \quad \text{(say)}. \end{aligned}\right\} \tag{5}$$

The integral $h(Q)$ is the approximation to $H(Q)$ computed from the approximate probability integrals f and g and $\theta(Q)$ the error thereby committed.

4. To estimate this error we have

$$\theta(Q) = \int_0^\infty \left\{\eta'(y) \int_0^{\psi(Q,y)} F'(x)\, dx\right\} dy + \int_0^\infty \left\{g'(y) \int_0^{\psi(Q,y)} \epsilon'(x)\, dx\right\} dy.$$

Applying partial integration to the first term, ordinary integration to the second term and noting (3), we reach

$$\theta(Q) = -\int_0^\infty \left\{F'(\psi(Q,y))\, \eta(y) \frac{\partial \psi}{\partial y}(Q,y)\right\} dy + \int_0^\infty \{g'(y)\, \epsilon(\psi(Q,y))\}\, dy$$

or

$$\theta(Q) = -\int_{\psi(Q,\infty)}^{\psi(Q,\infty)} F'(\psi)\, \eta(\chi(Q,\psi))\, d\psi + \int_0^\infty \{g'(y)\, \epsilon(\psi(Q,y))\}\, dy.$$

Now since $F' \geqslant 0$ and $g' \geqslant 0$, and since

$$\left|\int_{\psi(Q,0)}^{\psi(Q,\infty)} F'(\psi)\, d\psi\right| \leqslant \left|\int_0^\infty F'(\psi)\, d\psi\right| = 1, \quad \int_0^\infty g'(y)\, dy = 1,$$

we obtain immediately $\quad |\theta(Q)| \leqslant \max|\eta(y)| + \max|\epsilon(x)|,$

and this inequality proves the following lemma:

5. Lemma. Let x and y be independent variates with probability integrals $F(X)$ and $G(Y)$ respectively; let $f(X)$ and $g(Y)$ be approximations to F and G with errors $\epsilon(X)$ and $\eta(Y)$ respectively; finally, let $q = \phi(x, y)$ be a function of the variates x and y satisfying the above conditions (see § 2). If, then, the probability integral of q is computed from f and g, the error thereby committed is smaller than $\max|\epsilon| + \max|\eta|$. By repeated application of the lemma, the generalization to functions of more than two variates, $q = \phi(x_1, x_2, \ldots x_m)$, is obvious.

Correlations between χ^2 cells

By F. N. DAVID

1. The population studied is assumed to fall into k groups or strata, there being a proportion p_i $(i = 1, 2, \ldots, k)$ in the ith group. A sample of size N is randomly and independently drawn from the population, the number falling into the ith group being n_i. We write

$$m_i = Np_i \quad \text{and} \quad x_i = n_i - m_i.$$

It is well known that if the only restriction which is placed on the sample is that the totals of observation and expectation are made to agree, i.e. if

$$\sum_{i=1}^{k} x_i = 0,$$

then, writing r_{ij} for the coefficient of correlation between x_i and x_j,

$$r_{ij} = -\left(\frac{m_i m_j}{(N-m_i)(N-m_j)}\right)^{\frac{1}{2}}.$$

2. The coefficient of correlation between x_i and x_j when more than one restriction is placed on the sample is not easily determined. It may, however, be deduced from a consideration of the multivariate normal surface. The multivariate normal surface may be written

$$p(x_1 x_2 \ldots x_n) = \frac{1}{(2\pi)^{\frac{1}{2}n}\, \sigma_1 \sigma_2 \ldots \sigma_n R^{\frac{1}{2}}} \exp\left\{-\frac{1}{2R}\left[\sum_{i=1}^{n} R_{ii} \frac{x_i^2}{\sigma_i^2} + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} R_{ij} \frac{x_i x_j}{\sigma_i \sigma_j}\right]\right\},$$

where $R = \begin{vmatrix} 1 & r_{12} & r_{13} & \dots & r_{1n} \\ r_{21} & 1 & r_{23} & \dots & r_{2n} \\ r_{31} & r_{32} & 1 & \dots & r_{3n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{n1} & r_{n2} & r_{n3} & \dots & 1 \end{vmatrix}$,

R_{ij} is the minor obtained by omitting the ith row and the jth column, and r_{ij} is the coefficient of correlation between x_i and x_j. Consider the exponent, and write it as the sum of n linear squares, viz.:

$$(\alpha_{11}x_1+\alpha_{12}x_2+\alpha_{13}x_3+\dots+\alpha_{1n}x_n)^2+(\alpha_{22}x_2+\alpha_{23}x_3+\dots+\alpha_{2n}x_n)^2+\dots+\alpha_{nn}^2x_n^2.$$

Make the substitutions

$$\begin{aligned} z_1 &= \alpha_{11}x_1+\alpha_{12}x_2+\dots+\alpha_{1n}x_n, \\ z_2 &= \quad\ \alpha_{22}x_2+\dots+\alpha_{2n}x_n, \\ &\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ z_n &= \quad\ \alpha_{nn}x_n, \end{aligned}$$

and solve for the x's. We have then a series of relations of the form

$$\begin{aligned} x_1 &= \beta_{11}z_1+\beta_{12}z_2+\dots+\beta_{1n}z_n, \\ x_2 &= \quad\ \beta_{22}z_2+\dots+\beta_{2n}z_n, \\ &\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ x_n &= \quad\ \beta_{nn}z_n. \end{aligned}$$

If we now put $x_1 = x_2 = \dots = x_n = 0$, we have n equations which may be regarded as n planes in an n-dimensioned space. If θ_{ij} is the angle between the ith and the jth plane, then

$$\theta_{ij} = \cos^{-1}r_{ij}.$$

3. I believe this last relation to be well known,* although I have not been able to find a reference to its proof anywhere. The proof is, however, quite straightforward, making use of the well-known relations between the minors of a determinant. It is suggested that this result may be used to find the correlation between x_i and x_j of § 1. It is assumed that the number, N, in the sample, and the number of groups are large enough for the assumptions regarding χ^2 to be satisfied; that is to say, it is assumed that x_i is normally distributed about m_i for $(i = 1, 2, \dots, k)$. We consider a χ^2 of k groups where we have placed p linear restraints on the x's. The probability $P\{\chi^2 > \chi_0^2\}$, where χ_0^2 is some constant, is given by the multiple integral in a $(k-p)$-dimensioned space taken over the domain D defined by $\chi^2 > \chi_0^2$. We have then

$$P\{\chi^2 > \chi_0^2\} = \text{constant} \times \int\int_D \dots \int \exp\left[-\tfrac{1}{2}\sum_{i=1}^{k}\frac{x_i^2}{m_i}\right] dx_1 dx_2 \dots dx_{k-p}, \qquad (1)$$

where $x_{k-p+1}, \dots, x_k$ can be expressed in terms of $x_1, x_2, \dots, x_{k-p}$ by solving the equations by means of which the linear restraints are expressed. The expression under the integral sign is equivalent to a multivariate normal distribution in $k-p$ dimensions; accordingly, by writing the exponent in the way described in § 2 and finding the angle between the appropriate planes $x_i = 0$ and $x_j = 0$, r_{ij} can be deduced.

4. As an illustration consider the case of a population which is divided into five groups. A sample of N is drawn, and, using the notation of § 1, we have that the correlation between x_1 and x_2 is, for the case of one restraint,

$$r_{12} = -\left(\frac{m_1 m_2}{(N-m_1)(N-m_2)}\right)^{\frac{1}{2}}.$$

We begin by showing that this may be deduced by the method outlined here. If the one restraint is that the sum of the x's is zero, then we have

$$x_1+x_2+x_3+x_4+x_5 = 0 \quad \text{or} \quad -x_5 = x_1+x_2+x_3+x_4.$$

The exponent under the χ^2 integral may therefore be written (dropping the multiplier $-\frac{1}{2}$ for convenience)

$$\frac{x_1^2}{m_1}+\frac{x_2^2}{m_2}+\frac{x_3^2}{m_3}+\frac{x_4^2}{m_4}+\frac{(x_1+x_2+x_3+x_4)^2}{m_5}.$$

---

* It is easily deduced from determinantal relations given by K. Pearson in lectures and in various papers dealing with multiple correlation. M. G. Kendall points out that it follows as a natural result of the discussion on p. 372 of his *Advanced Theory of Statistics*, vol. 1. Sheppard used the bivariate result as a means of estimating the correlation coefficient.

This may be rewritten as

$$\left[x_1\left(\frac{m_1+m_5}{m_1m_5}\right)^{\frac{1}{2}}+x_2\left(\frac{m_1}{m_5(m_1+m_5)}\right)^{\frac{1}{2}}+x_3\left(\frac{m_1}{m_5(m_1+m_5)}\right)^{\frac{1}{2}}+x_4\left(\frac{m_1}{m_5(m_1+m_5)}\right)^{\frac{1}{2}}\right]^2$$
$$+\left[x_2\left(\frac{m_1+m_2+m_5}{m_2(m_1+m_5)}\right)^{\frac{1}{2}}+x_3\left(\frac{m_2}{(m_1+m_2+m_5)(m_1+m_5)}\right)^{\frac{1}{2}}+x_4\left(\frac{m_2}{(m_1+m_2+m_5)(m_1+m_5)}\right)^{\frac{1}{2}}\right]^2$$
$$+\left[x_3\left(\frac{N-m_4}{m_3(m_1+m_2+m_5)}\right)^{\frac{1}{2}}+x_4\left(\frac{m_3}{(N-m_4)(m_1+m_2+m_5)}\right)^{\frac{1}{2}}\right]^2+\left[x_4\left(\frac{N}{m_4(N-m_4)}\right)^{\frac{1}{2}}\right]^2.$$

Substituting z_1 for the expression in the first bracket, z_2 for the second and so on, and solving for x_1, x_2, x_3 and x_4 in terms of the z's, we have

$$x_1=\left(\frac{m_1m_5}{m_1+m_5}\right)^{\frac{1}{2}}z_1-\frac{m_1m_2^{\frac{1}{2}}}{(m_1+m_5)^{\frac{1}{2}}(m_1+m_2+m_5)^{\frac{1}{2}}}z_2-\frac{m_1m_3^{\frac{1}{2}}}{(m_1+m_2+m_5)^{\frac{1}{2}}(N-m_4)^{\frac{1}{2}}}z_3-\frac{m_1m_4^{\frac{1}{2}}}{N^{\frac{1}{2}}(N-m_4)^{\frac{1}{2}}}z_4,$$

$$x_2=\frac{(m_1+m_5)^{\frac{1}{2}}m_2^{\frac{1}{2}}}{(m_1+m_2+m_5)^{\frac{1}{2}}}z_2-\frac{m_2m_3^{\frac{1}{2}}}{(m_1+m_2+m_5)^{\frac{1}{2}}(N-m_4)^{\frac{1}{2}}}z_3-\frac{m_2m_4^{\frac{1}{2}}}{N^{\frac{1}{2}}(N-m_4)^{\frac{1}{2}}}z_4,$$

$$x_3=\frac{(m_1+m_2+m_5)^{\frac{1}{2}}m_3^{\frac{1}{2}}}{(N-m_4)^{\frac{1}{2}}}z_3-\frac{m_3m_4^{\frac{1}{2}}}{(N-m_4)^{\frac{1}{2}}N^{\frac{1}{2}}}z_4,$$

$$x_4=\frac{(N-m_4)^{\frac{1}{2}}m_4^{\frac{1}{2}}}{N^{\frac{1}{2}}}z_4.$$

The angle between the planes $x_1=0$ and $x_2=0$ will be

$$\theta_{12}=\cos^{-1}-\frac{m_1m_2}{N}\Big/\left(\frac{m_1(N-m_1)}{N}\right)^{\frac{1}{2}}\left(\frac{m_2(N-m_2)}{N}\right)^{\frac{1}{2}}=\cos^{-1}-\left(\frac{m_1m_2}{(N-m_1)(N-m_2)}\right)^{\frac{1}{2}}.$$

For this case the result may also be reached by calculating

$$\theta_{34}=\cos^{-1}-\left(\frac{m_3m_4}{(N-m_3)(N-m_4)}\right)^{\frac{1}{2}}$$

and deducing θ_{12} by symmetry.

5. It will be noted that the restraint placed on χ^2 in the preceding section is equivalent to making the totals of observation and expectation agree, as was pointed out in § 1. We now proceed to place such further restraints on χ^2 as are made when moments of observation and expectation are put in agreement. Admittedly this narrows the field of investigation to a certain extent, but it may be argued that such restraints are those which are most often met in practice. We assume therefore that the two restraints placed on the sample are

$$x_1+x_2+x_3+x_4+x_5=0,$$
$$-2x_1-x_2+x_4+2x_5=0,$$

or that the totals of observation and expectation have been made to agree, and that the population mean has been estimated from the sample. We solve for x_4 and x_5, express the quantity

$$\sum_{i=1}^{5}\frac{x_i^2}{m_i}$$

in terms of x_1, x_2 and x_3 only, write the expression so obtained as the sum of three linear squares and, after substitution of the new variables z_1, z_2 and z_3, solve for the x's. We thus obtain

$$x_1=\frac{(m_1m_4m_5)^{\frac{1}{2}}}{A^{\frac{1}{2}}}z_1-\frac{m_1m_2^{\frac{1}{2}}(12m_5+6m_4)}{A^{\frac{1}{2}}B^{\frac{1}{2}}}z_2-\frac{m_1m_3(8m_5+3m_4-m_2)}{B^{\frac{1}{2}}C^{\frac{1}{2}}}z_3,$$

$$x_2=\frac{A^{\frac{1}{2}}m_2^{\frac{1}{2}}}{B^{\frac{1}{2}}}z_2-\frac{m_2m_3^{\frac{1}{2}}(6m_5+2m_4+2m_1)}{B^{\frac{1}{2}}C^{\frac{1}{2}}}z_3,$$

$$x_3=\frac{B^{\frac{1}{2}}m_3^{\frac{1}{2}}}{C^{\frac{1}{2}}}z_3,$$

where

$$A=m_4m_5+16m_1m_5+9m_1m_4,$$
$$B=A+m_2(9m_5+4m_4+m_1),$$
$$C=B+m_3(4m_5+m_4+m_2+4m_1).$$

The cosine of the angle between the planes $x_1=0$, $x_2=0$ is

$$r_{12}=-\frac{(12m_5+6m_4+2m_3)(m_1m_2)^{\frac{1}{2}}}{F^{\frac{1}{2}}G^{\frac{1}{2}}},$$

where
$$F = m_5m_4 + 4m_5m_3 + 16m_5m_1 + m_4m_3 + 9m_4m_1 + 4m_3m_1,$$
$$G = m_5m_4 + 4m_5m_3 + 9m_5m_2 + m_4m_3 + 4m_4m_2 + m_3m_2.$$

Again it will be noted that owing to symmetry in the restraints r_{43} may be deduced from r_{12} by the substitution of the appropriate indices.

6. It is seen, as in fact is expected, that the correlation between x_i and x_j for more than one restraint depends on the shape of the population from which the sample has been drawn, i.e. it depends on the relative values of $m_1, m_2, \ldots$, and it will also depend on the type of restraint which is placed on the sample. The behaviour of these correlations is interesting and study of them arose out of a larger investigation into the relations between the signs of the deviations as the number of moment-restraints is increased. Thus it is found that the correlation between x_1 (say) and x_2, negative for one restraint, increases numerically to -1 as the number of moment-restraints increase. On the other hand, the correlation between x_1 and x_3, negative for one restraint, decreases numerically and finally increases to $+1$ with increasing numbers of moment-restraints. Generally we may expect the correlations between two x's with odd subscripts or two x's with the even subscripts to tend to $+1$, and the correlations between an x with an odd subscript and an x with an even subscript to tend to -1.

7. Several tests regarding the signs of the deviations, x, have been proposed recently and the opinion has been expressed that these signs could be regarded as effectively independent for the case where more than one restraint is placed on the material. For such sign tests what is important is the correlation between adjacent deviations and this will, for the case of moment-restraints at any rate, increase numerically to -1; for the case of extreme deviations this increase will be a rapid one. For illustrative purposes I have considered two cases, the first where all the m's are given equal weight, i.e. letting $m_1 = m_2 = m_3 = \ldots = m_k$ and second where $m_1 = m_k < m_2 = m_{k-1} < m_3 = m_{k-2}$* etc. For the first case the correlation between x_1 and x_2 was found for five groups and 1, 2, 3 moment-restraints, seven groups and 1, 2, 3, 4 moment-restraints, nine groups and 1, 2, 3, 4 moment-restraints. The correlation coefficients are given in Table 1.

Table 1. *Correlation between x_1 and x_2; equal weighting of expectations*

No. of groups	5			7				9			
No. of restraints	1	2	3	1	2	3	4	1	2	3	4
Correlation, r_{12}	−0·25	−0·76	−0·92	−0·17	−0·58	−0·87	−0·95	−0·12	−0·43	0·80	−0·92

For the second case the correlations were compared for five groups only and the results are given in the last row of Table 2.

Table 2. *Correlation between x_1 and x_2; expectations weighted equally and unequally*

No. of groups	5		
No. of restraints	1	2	3
Correlation, r_{12}: Equal weights	−0·25	−0·76	−0·92
Unequal weights	−0·12	−0·57	−0·94

The precise numerical value of these correlation coefficients is not important; what is important is the rapidity with which the correlation increases as the number of moment-restraints is increased. This rapid increase would suggest that any test derived on the assumption of the independence of signs of deviations, for extreme observations at any rate, is of doubtful validity.

8. As a check on theory I considered the first 100 samples of the 208 samples, each of 200 observations, used by Neyman & Pearson (1928) as an illustration in their χ^2 paper. The population in this case is a

* Actually the frequencies were obtained by dividing the normal curve between mean $\pm 3\sigma$ into groups with equal base range.

cubic divided into eight groups. I have shown previously (David, 1947) that for the case where one restraint is placed on the material the order of the signs of the deviations can be regarded as random, thus indicating that the effect of the correlation is not felt. We now consider the case when four restraints are placed on the material; the totals of expectation and observation are made to agree and the first three moments are estimated from the data. The correlation between x_1 and x_2 is now equal to $-0{\cdot}840$, and the correlation between x_1 and x_3 is $+0{\cdot}213$. It will be noted that r_{12} is comparable in magnitude with the correlations given in Table 1 for 4 restraints and equal weighting of the expectations. It is clear that both these correlations are too large to be neglected, that between x_1 and x_2 being of sufficient magnitude seriously to invalidate any sign test which is based on a hypothesis of independence or of randomness.

9. It will be realized that the manner in which restraints are imposed on the differences $x_i = n_i - m_i$ is not quite the same in the problem we have been considering as in that arising in tests where a theoretical law is fitted to the data. In the former case the x_i satisfy certain linear relations because we confine attention to the set of samples for which the n_i satisfy certain conditions. In the second case the n_i are unrestricted except for the condition

$$\sum_{i=1}^{k} n_i = N,$$

but in calculating x_i we substitute $\hat{m}_i$ for m_i, where these estimates are so chosen that $n_i - \hat{m}_i$ satisfy similar conditions.

This question has been discussed at some length only by Neyman & Pearson (1928) and is invariably inadequately treated in statistical text-books. I do not propose to discuss the matter fully here but would offer certain remarks which lead me to believe that the substitution of sample estimates, $\hat{m}$, for the population values, m, in (1) will not seriously invalidate the calculated correlations.

We may consider for simplicity a population of four groups. As before, we shall have

$$x_i = n_i - m_i,$$

and we shall write

$$X_i = \hat{m}_i - m_i = f_i(\phi),$$

where ϕ is the estimate of the population parameter which is calculated from the observations. Because ϕ is an estimate it will vary from sample to sample and the estimated values X_1, X_2, X_3, X_4 will lie on a curve (termed by Neyman & Pearson the population locus), which will depend on ϕ alone. It is clear therefore that for any given set of observations x_1, x_2, x_3, x_4 the estimated population point will depend on the method of fitting used, or, perhaps more precisely, on the method of estimation of ϕ.

If the structure of the restraint placed on the observations for the purpose of estimating ϕ is such as to lead to a minimum

$$\chi^2 = \sum_{i=1}^{4} \frac{(n_i - \hat{m}_i)^2}{\hat{m}_i},$$

then the correlations, obtained under the assumption that (1) is true when $\hat{m}_i$ is substituted for m_i, will be approximately correct. In general the method of moments will not lead to this minimum χ^2 exactly, but the fitted expectations will not usually be very different from the minimum values, and the error made in assuming (1) is true will not be large.

REFERENCES

David, F. N. (1947). *Biometrika*, **34**, 299.
Neyman, J. & Pearson, E. S. (1928). *Biometrika*, **20A**, 274.

Note on 'Proofs of the distribution law of the second order moment statistics'

By JOHN WISHART

Since the paper under the above title was published (1948, *Biometrika*, **35**, 55), my attention has been called to a paper by E. Sverdrup, 'Derivation of the Wishart Distribution of the Second Order Sample Moments by straightforward Integration of a Multiple Integral' (1947, *Skand. AktuarTidskr.* **30**, 151). A copy has also been received of a paper by R. D. Narain, 'A New Approach to Sampling Distributions of the Multivariate Normal Theory', in which the same distribution is derived. This paper will shortly be published in India.

REVIEW

Probit Analysis. By D. J. Finney, M.A. Cambridge University Press. 1947. Price: 18*s*.

The publication of books relating to specialized statistical techniques is a welcome feature of recent years. In such books it is possible to give an account sufficiently detailed to provide a practical guide to the user of these techniques in most of the problems to be encountered, as opposed to the necessarily more restricted treatment of any particular topic in a general text-book. Mr Finney's volume on probit analysis, with special emphasis on applications in biological research, is an excellent example of this specialized type of book, not only taking full advantage of the possibility of exhaustive treatment, but also presenting the arguments in such a way as to ease the reader's task in mastering the methods described.

Starting from a simple description of the types of problem to be treated, techniques covering almost every eventuality in the field commonly described as 'dosage-mortality' are developed and their practical application illustrated by means of numerical examples. The standard methods of probit analysis are carefully described and very useful sets of tables (Tables I–IV) are provided. Tables III and IV, in particular, should considerably facilitate the computation of working probits. Table III gives both maximum and minimum working probits for expected probits at intervals of 0·1; Table IV enables the working probit to be obtained directly from the provisional probit and the percentage 'kill'. Table II, apart from giving values of Q/Z, gives values of the weighting coefficient allowing for the effect of natural mortality. Chapter 6 is devoted to the consideration of methods of allowing for natural mortality and should prove most useful. While the methods described are not new, it is the first time that they have appeared in a systematic treatise in such a way as to bring them to general notice. Similar remarks might be made about the other special features of the book—the treatment of factorial experiments, of the joint action of different poisons and of quantitative responses, for example. Most of the methods described have already been published in individual papers, but Mr Finney has performed an important task in bringing them together in one consolidated account. The method of development is not, of course, theoretically complete or rigorous, but it is logically clear, and should provide an invaluable aid to experimentalists in their appreciation of the statistical tests described.

It may seem ungenerous to extract for special criticism the very few passages in this useful book which seem to be ambiguous or capable of improvement. Such criticism will be made, however, as a modest attempt to help readers of the work to avoid certain difficulties and confusion which they might possibly experience. It appears to the reviewer that some difficulties may result from the author's desire to simplify the presentation of the underlying theory. While it must be admitted that some simplification is necessary, this should not be such as to allow of false impressions of the theoretical situation. As it appears that considerable knowledge of general statistical theory is expected of the reader, there would seem to be no need to avoid the use of simple technicalities in clarifying the exposition.

We may instance the arguments for introducing the heterogeneity factor (p. 33), the working probit (p. 48) and the correction for natural mortality (p. 88). In the first of these cases it should be noted that the interval of estimation will be on the average too long if there is no heterogeneity; while if heterogeneity is present the occasional use of normal factors will be incorrect. On p. 48, the implication that it is merely the asymmetry of the distribution of p which gives rise to the need for working probits gives a rather incomplete notion of the theory. Finally, the arguments on corrections for natural mortality could be much simplified if they were based directly on equation (6·2).

The use of both LD 50 and ED 50 to symbolize the same population parameter, LD 50 being restricted to cases where the response is death, may be somewhat confusing. We note, indeed, that the author himself uses the symbol ED 50 on p. 124, when considering 'kills'. It may save some confusion to point out that the quantity x_1x_2, defined as x_2 on p. 117, is afterwards represented as x_{12} (there is also a misplaced bracket in the first equation on this page); there is also a misprint on p. 89, where the reference to equation (3·5) should be to equation (3·4).

The comparatively minor importance of the above criticisms must be emphasized. The book as a whole is a valuable addition to the statistician's library and is indispensable for persons concerned with 'dosage-mortality' problems, in whatever form they arise. Apart from its other merits, the very full bibliography makes the book invaluable as a reference index in its own field. May other works of this specialized nature maintain the standard set by Mr Finney!

N. L. J.

CORRIGENDA

1. Moments of the mean deviation from the mean in samples from a normal population. (See the reference in paper by H. J. Godwin, p. 308 above.) The following corrections should be made:

(1·1) R. C. Geary (1936), *Biometrika*, **28**:

p. 300 In equation (22) the coefficient of n^{-4} in the expansion for m_4'' should read

$$(51-352a^2+\tfrac{2136}{5}a^4), \quad \textit{not} \quad (51-352a^2+427a^4).$$

p. 301 In equation (24) the coefficient of n^{-4} in the expression for $m_4''-m_4'$ should read

$$0{\cdot}11463500, \quad \textit{not} \quad 0{\cdot}03357805.$$

(1.2) E. S. Pearson (1945), *Biometrika*, **33**, p. 252:

In equation (4) the coefficient of ν^{-4} in the expansion for λ_4 should read $-0{\cdot}038946$, *not* $-0{\cdot}120003$. This correction makes alterations in the values for λ_4 and β_2 given in the table of moments of the mean deviation on p. 253, which should read:

Sample size n	λ_4	β_2
4	0·001 963	3·252
5	·000 9912	3·197
6	·000 5672	3·161
8	·000 2356	3·118
10	·000 1195	3·093
12	·000 06868	3·076
15	·000 03493	3·061
20	·000 01464	3·045

2. *Biometrika*, **35**, 181 (1948):

In the list of references at the end of F. Yates's paper, that to E. S. Pearson (1923) should read *Biometrika*, **14**, 261, *not Biometrika*, **7**, 248.

(*All Rights reserved*)

BIOMETRIKA. Vol. XXXV, Parts III and IV

CONTENTS

	PAGE
Some further notes on the use of matrices in population mathematics. By P. H. LESLIE	213–245
The transformation of Poisson, binomial and negative-binomial data. By F. J. ANSCOMBE	246–254
Some theorems on time series. II. The significance of the serial correlation coefficient. By P. A. P. MORAN	255–260
Some results in the testing of serial correlation coefficients. By M. H. QUENOUILLE	261–267
Fractional replication arrangements for factorial experiments with factors at two levels. By K. A. BROWNLEE, B. K. KELLY and P. K. LORAINE	268–276
The relationship between finite groups and completely orthogonal squares, cubes and hypercubes. By K. A. BROWNLEE and P. K. LORAINE	277–282
Systematic sampling of continuous parameter populations. By A. E. JONES	283–290
Continuation of Dr Jones's paper. By M. G. KENDALL	291–296
The precision of observed values of small frequencies. By J. B. S. HALDANE	297–300
Note on Professor Haldane's paper regarding the treatment of rare events. By E. S. PEARSON	301–303
A further note on the mean deviation. By H. J. GODWIN	304–309
A note on the asymptotic distribution of range. By D. R. COX	310–315
On the role of variable generation time in the development of a stochastic birth process. By DAVID G. KENDALL	316–330
2 × 2 Tables; the power function of the test on a randomized experiment. By E. S. PEARSON and MAXINE MERRINGTON	331–345
Statistical analysis of a non-orthogonal tri-factorial experiment. By W. L. STEVENS	346–367
Comparisons of heights and weights of German civilians recorded in 1946–7 and Royal Air Force and other British series. By G. M. MORANT	368–396
Testing the significance of correlation between time series. By G. H. ORCUTT and S. F. JAMES	397–413
MISCELLANEA	
Note on the median of a multivariate distribution. By J. B. S. HALDANE	414–415
A property of rank correlations. By H. E. DANIELS	416–417
Approximation errors in distributions of independent variates. By H. O. HARTLEY	417–418
Correlations between χ^2 cells. By F. N. DAVID	418–422
Note on 'Proofs of the distribution law of the second order moment statistics.' By JOHN WISHART	422
REVIEW	
Probit Analysis	423

A volume of *Biometrika* containing about 400 pages, with plates and tables, will be published annually in two half-yearly issues.

Papers for publication should either be sent to

PROFESSOR E. S. PEARSON, Department of Statistics, University College, London, W.C. 1,

or if more convenient to

DR JOHN WISHART, School of Agriculture, Cambridge.

M. G. KENDALL, Esq., 9 Duke's Lodge, Holland Park, London, W. 11.

It is a condition of publication in *Biometrika* that the paper shall not already have been issued elsewhere, and will not be reprinted without leave of the Editors.

Contributors receive 25 copies of their papers free. Joint authors 15 copies each.

The subscription price, *payable in advance*, is Inland 45*s. net* per volume and Abroad 54*s. net* (including packing and postage). Owing to the scarcity of early volumes, the following rates must now be charged for complete sets. Vols. I—XXXIV, including XX^S: £133 in wrappers, not including postage. At present certain volumes are out of print, but steps are being taken to re-issue these as quickly as printing facilities permit. Recent volumes may still be obtained at the wrapper price; this is 64*s.* inland, including postage. Index to Vols. I to V, 2*s. net*. Index to Vols. I to XV, 5*s. net*. Cheques must be made payable to *Biometrika*, crossed "*a/c Biometrika Trust*" and sent to The Secretary, *Biometrika Office*, Department of Statistics, University College, London, W.C. 1, to whom all orders for series, single copies and offprints should be addressed. All foreign cheques must be drawn in sterling and on a Bank having a London Agency.

First printed in Great Britain at the University Press, Cambridge

Reprinted by offset-litho by Percy Lund, Humphries & Co., Ltd, Bradford